2019

Identifying College Students' Course-Taking Patterns In Stem Fields

Fahimeh Bahrami
University of Vermont

IDENTIFYING COLLEGE STUDENTS' COURSE-TAKING PATTERNS IN STEM FIELDS

A Dissertation Presented

by

Fahimeh Bahrami

to

The Faculty of the Graduate College

of

The University of Vermont

In Partial Fulfillment of the Requirements
for the Degree of Doctor of Philosophy
Specializing in Educational Leadership and Policy Studies

May, 2019

Defense Date: March 28, 2019
Dissertation Examination Committee:

Tammy G. Kolbe, Ed.D., Advisor
Cynthia J. Forehand, Ph.D., Dean of the Graduate College, Chairperson
Scott. R. Thomas, Ph.D.
Sean M. Hurley, Ph.D.
Maureen Neumann, Ph.D.
Alexander, Yin, Ph.D.

ABSTRACT

In spite of substantial investments in science, technology, engineering, and mathematics (STEM) education, low enrollment and high attrition rate among students in these fields remain an unmitigated challenge for higher education institutions. In particular, underrepresentation of women and minority students with STEM-related college degrees replicates itself in the makeup of the workforce, adding another layer to the challenge. While most studies examine the relationship between student characteristics and their outcomes, in this study, I take a new approach to understand academic pathways as a dynamic process of student curricular experiences that influence his/her decision about subsequent course-takings and major field of the study. I leverage data mining techniques to examine the processes leading to degree completion in STEM fields. Specifically, I apply Sequential Pattern Mining and Sequential Clustering to student transcript data from a four-year university to identify frequent academic major trajectories and also the most frequent course-taking patterns in STEM fields. I also investigate whether there are any significant differences between male and female students' academic major and course-taking patterns in these fields.

The findings suggest that non-STEM majoring paths are the most frequent academic pattern among students, followed by life science trajectories. Engineering and other hard science trajectories are much less frequent. The frequency of all STEM trajectories, however, declines over time as students switch to non-STEM majors. The switching rate from non-STEM to STEM fields overtime is, however, much lower. I also find that male and female students follow different academic pathways, and these gender-based differences are even more significant within STEM fields.

Students' course-taking patterns also suggest that taking engineering and computer science courses is predominantly a male course-taking behavior, while females are more likely to pursue academic pathways in life science. I also find that STEM introductory courses - particularly Calculus I, Calculus II and Chemistry I – are gateway courses, that serve as potential barriers to pursuing degrees in STEM-related fields for a large number of students who showed an initial interest in STEM courses. Female students were more likely to switch to non-STEM fields after taking these courses, while male students were more likely to drop out of college overall.

In addition to the study's findings on students' academic pathways toward attaining a college degree in a STEM-related field, this study also shows how data mining techniques that leverage data about the sequence of courses students take can be used by higher education leaders and researchers to better understand students' academic progress and explore how students navigate and interact with college curriculum. In particular, this study demonstrates how these analytic approaches might be used to design and structure more effective course taking pathways and develop interventions to improve student retention in STEM fields

ACKNOWLEDGMENT

I would like to express my gratitude to a number of people who have supported me throughout the dissertation process. First, my dissertation advisor Dr. Tammy Kolbe who was a consistent source of support throughout this process. She is an inspiration and a role model who constantly pushed me, never allowing me to settle for less than what she thought I was capable of. I could not ask for a better advisor and mentor. I would also like to thank the rest of my thesis committee: Dean Scott Thomas, Dr. Sean Hurley, Dr. Maureen Neumann, and Dr. Alex Yin for their encouragement throughout this process and for asking the hard and crucial questions that broadened my viewpoint and made me think about my research from various perspectives.

My sincere thanks also go to Dr. Alex Yin and Larry Granillo from Office of Institutional Research who provided the necessary data for this research. Without their cooperation and support this project could have not been conducted.

I am very lucky to have had the friendship of Sarah Maria Childs throughout my doctoral program. I owe countless moments of joy that made the arduous process of dissertation writing possible to her. Her friendship is one of things I value most about my grad school experience.

Last but not the least, I would like to thank my family. Ata, thanks for putting up with me during these stressful and difficult years of study and providing me with nothing but love and encouragement. I would have never been able to accomplish this without your support and love. Narges and Esra, my lovely girls- you have been very

understanding and patient with this long process and lifted up my spirit with your love whenever I felt down.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1: INTRODUCTION

Higher education institutions in the United States face a serious challenge in attracting and retaining students in Science, Technology, Engineering, and Mathematics (STEM) fields. Evidence shows that the number of college students who intend to pursue a degree in STEM fields has been consistently lower than other fields (Hill, Corbett, & St. Rose, 2010). Only 15 percent of freshmen students enrolled in the U.S. post-secondary education in 2011-12 reported that they intended to declare a major in a STEM fields (National Science Board, 2016). An additional concern is that roughly half of those undergraduates who show an initial interest in a STEM-related major in college switch out of these fields within their first two years of study, and very few students who were initially non-STEM majors switch to STEM majors (Chen, 2013; Kokkelenberg & Sinha, 2010). Although low completion and switching rates are not unique to STEM fields, it is more concerning in these fields because many STEM leavers are actually high-performing students who might make valuable additions to STEM workforce (Chen, 2015; Seymour, 2002).

In addition to low enrollment and persistence rates, there are significant gender and racial gaps in STEM fields - both in terms of the individuals who intend to enroll in these fields and those who successfully finish degrees. Evidence suggests that women and underrepresented groups do not pursue or complete STEM-related degrees at higher rates than their counterparts (Bebe-vroman, Juniewicz, Lucarelli, Fox, Nguyen, & Tjang, 2017; Bowen, Chingos, & McPherson, 2009; Chen, 2013; George-Jackson, 2016; Hill et al., 2010; Huang, Taddese, & Walter, 2000; Seymour & Hewitt, 1997; Simpson, 2001).

This raises the question of whether higher education institutions are also capable of ensuring equal educational opportunities for all students.

Past studies that investigated college students' persistence in STEM fields focused primarily on individual and institutional characteristics and their impact on student outcomes, particularly existing disparities in enrollment and outcomes among different student groups. However, existing research has paid less attention to a student's academic behavior throughout college. There is no doubt that individual and institutional characteristics play important roles in determining a student's academic performance; however, such studies offer very little insight into the processes that lead to graduation or noncompletion within educational institutions. Understanding this process is an important consideration when evaluating differences in student outcomes. A student's pathway toward a degree is a dynamic process of curricular experiences that influence his/her decision about subsequent course-takings and major field of the study (Chen, 2013; Shapiro & Sax, 2011). Yet, most existing research subscribes to a traditional input/output conceptual framing of the problem, and likewise employs analytic approaches that describe the relationship between some input-related variables and whether students persist toward or complete degrees in STEM fields. In fairness, the rationale for this framing and analytic approach is due to the fact that even the most detailed of linear modeling techniques do not have the capacity to describe the dynamic processes at work that shape the various academic trajectories students take to earn a degree. Therefore, our knowledge about what actually happens along students' academic pathways through STEM pipeline is very limited. That is, while students' coded transcript data detailing their progress toward a degree is collected by higher education institutions, researchers have rarely considered using these data to identify the pathways that align with academic major selection and successful degree attainment in STEM fields.

Studying students' academic pathways – particularly their course taking behavior while in college – could provide valuable insight into the phases of study or sequence of courses that comprise students' experiences. This information can then be used to answer questions of how and why students decide to persist toward and complete degrees in STEM-related fields. In other words, by examining students' course taking patterns, we could potentially identify courses that function as a road block for different student groups pursuing STEM fields, and conversely, identify the paths students take toward successfully completing a degree in a STEM field. These patterns may also offer valuable information about how students' academic pathways are related to decisions to leave or switching fields within STEM majors. For instance, evidence suggests that there may be gender differences among STEM fields (George-Jackson, 2016; Kokkelenberg & Sinha, 2011; Ost, 2010) – e.g., biological/life science attract more female students than hard sciences such as physics, engineering, and computer sciences. Investigating differences in course-taking patterns may provide a better understanding of such gender-based patterns within STEM fields as well (Kokkelenberg & Sinha, 2011).

Despite all the potential benefits, students' academic behavior including major and course-taking patterns in educational institutions, in general, and in STEM fields in four-year colleges, in particular, has been rarely examined by researchers. This may be due to the challenges with empirically mapping these patterns. However, recently, significant progress has been made in different fields to develop and apply new methodologies to discover useful patterns in student course taking. These methodologies, generally referred to as “Data Mining” are devoted to extracting hidden knowledge from vast amounts of daily accumulated data. In the field of education, the application of such methods has been mostly limited to E-learning, but rarely applied to traditional educational settings (Luan, 2002). A rare exception is the few research projects (e.g., Crosta, 2014; Wang, 2016) conducted in community college settings where the researches have taken an innovative approach and used data mining techniques to understand student course taking patterns.

This study addresses limitations in current research by applying data mining techniques to better understand students’ academic major and course taking patterns in STEM fields. Identifying these patterns may not only shed light on course taking paths that lead to STEM major selection and ultimate degree attainment, but also identify particular types of courses or sequence of courses that may act as gatekeeper, leading some students to leave their field by switching to other fields or leaving a university altogether.

Background

Existing studies that investigate college students’ enrollment report a consistent low enrollment rate in undergraduate majors in STEM fields (Chen, 2013; Hill et al., 2010). Depending on the definition of STEM fields and undergraduate STEM majors, different enrollment rates have been reported by different studies (Chen, 2013). That said, the majority of researchers agree that enrollment rates in STEM-related academic majors are significantly lower than non-STEM fields. For instance, the National Science Board's report on Science and Engineering Indicators (2016) found that STEM majors accounted for just 20 percent of all undergraduate students enrolled in U.S. post-secondary education during 2011-12 academic year.

While STEM employment has grown at twice the rate of other non-STEM occupations and there are significant economic incentives (e.g., higher wages) for people to earn a STEM degree, we have not seen a solid increase in the number of students entering STEM fields (Bowen, Chingos, & McPherson, 2009). Lowell, Salzman and Bernstein (2009) examined six cohorts of students reaching back to the early 1970 using several longitudinal data sets. Their findings affirm that, on average, there have been no substantive changes in the proportion of high school graduates who enroll in STEM-related academic majors between 1972 to 2000. Their study also suggests that high-performing high school students are more likely to enter STEM fields than their low performing counterparts. The most concerning finding, however, is that there has been a rapid decline in the enrollment of top achievers in STEM fields from 28.7 percent in the 1992/97 cohort to 13.8 percent in the 2000/05 cohort.

In addition to the low enrollment rates in academic STEM majors, a number of studies report a gender difference in the share of students who pursue academic majors in a STEM-related field. In their longitudinal study, Xie and Shauman (2003) find a large gender imbalance among high school seniors intending to major in science and engineering in college. For every two males there was only one female who expressed interest in an academic major in a STEM-related field. Other studies (e.g., Chen, 2009;

Hill et al., 2010) portray a similar picture. According to annual American freshman record (Pryor et al., 2010), among first-year college students nationwide, only 17.3 percent of women report planning to major in a STEM field compared to 32.2 percent among men. Similarly, Chen (2013) reports a much higher percentage of STEM enrollment for men compared to women (around 33 percent vs. 14 percent), especially in engineering, physical sciences, and computer sciences. Other studies confirm the same results (e.g., George-Jackson, 2016; Simpson, 2001).

Another difficulty that institutions of higher education face is retaining students who initially intend to complete academic majors in STEM-related fields. The National Center for Education Statistics examined college students' paths into and out of STEM fields using several longitudinal data sets found striking results: Between 2003 and 2009, 48 percent of bachelor's degree students who pursued an academic major in a STEM-related left these fields by the spring 2009 (Chen, 2013). Of students who did not complete a STEM-related major, half switched their majors to a non-STEM field and the rest left postsecondary education without earning a degree. While switching majors is common among college students, other studies have found an even higher share of students switch out of STEM fields – as many as 50 percent (e.g., Kokkelenberg & Sinha, 2010). Moreover, completion outcomes vary within STEM fields. A higher rate of students in engineering and computer science leave the college without earning a degree compared to other STEM fields (Chen, 2013).

Many studies also find racial/ethnic disparities in persistence and attainment rate among students pursuing STEM-related academic majors (Bowen et al., 2009; Chen, 2015; Simpson, 2001). The disparity is mostly between Black and Hispanic students and their White counterparts. Among racial/ethnic groups, only Asian students have a higher persistence rate compared to White students (Bowen et al., 2009; Chen, 2015; George-Jackson, 2016; Huang, Taddese, & Walter, 2000; Simpson, 2001). Gender disparities in STEM persistence, however, have been a subject of debate. While some researchers (George-Jackson, 2016; Huang et al., 2000; Seymour & Hewitt, 1997) found a significant gap between male and female students in STEM degree completion, others did not find such a gap (e.g., Chen, 2013; Kokkelenberg & Sinha, 2011). Such divergent results might be explained as a result of differences across STEM fields in persistence toward degree. For example, while a larger percentage of men pursue and complete degrees in the hard sciences (e.g., physical sciences, engineering, and computer science), women have pursued and persisted toward degrees in life science at higher rates than men (George-Jackson, 2011). For example, Bebe-vroman et al. (2017) found that not only do smaller shares of female undergraduates plan to major in computer science than their male peers, they are also more likely to leave the major before receiving a degree. This has led some to argue that failing to account for differences between men and women in persistence patterns, particularly in the soft and hard sciences, can lead to misunderstanding gender disparities within STEM fields (George-Jackson, 2016; Kokkelenberg & Sinha, 2011; Ost, 2010).

To summarize, higher education institutions in the US face a real challenge in both enrolling and retaining college students in STEM-related academic majors. Moreover, there are disparities among different groups of students in both STEM major selection and degree attainment. Minorities and women are less likely to pursue degrees and more likely to leave such fields without earning a degree. Moreover, to date, most studies that examine enrollment and persistence rates in STEM fields have focused on descriptive analysis and have not paid attention to students' curricular experience and how that influences their subsequent course-taking pattern, major selection, and, finally, degree attainment.

Study Overview

The purpose of this study is to examine academic pathways through college among students who may be considering a STEM-related major. First, I describe the share of students with declared majors when they first matriculated as a degree seeking student at the university. I then explore how this initial distribution of students' academic majors' changes over time, and the extent to which patterns differ for female and male students. Specifically, I consider three research questions:

1) What major and course-taking patterns are aligned with degree attainment in STEM fields?
2) Are there any significant differences between men and women's academic major and course-taking patterns within STEM fields?
3) In which phase of their program of study do STEM students switch to other fields?

To answer these questions, I leverage recent developments of data mining techniques and apply these strategies to coded undergraduate transcript data at a four-year university. My approach to identifying student course-taking patterns is similar to techniques used in market basket analysis to identify costumer shopping behavior. Using Sequence Pattern Mining techniques, I create sequences of academic majors and STEM related courses a student takes each semester. Then, by clustering those sequences, I identify patterns of academic major and course-taking that are common to students who successfully pursue the fields and obtain a degree. Such patterns reveal which groups of students took similar academic major paths and also which groups of students decided to switch to other fields or dropped out of college after declaring a major in STEM fields. In terms of students' course-taking patterns, these methods reveal the sequences of STEM courses and the characteristics of student group that stopped taking STEM courses. As a result, I identify the so-called "gate-keeper courses" that compel certain groups of students not to take any further STEM courses, switching to another field or dropping out of the college. Identifying sequential patterns in student decisions about academic majors and their course-taking hold great promise for informing higher education policy and practice, particularly designing and structuring effective pathways to improve student retention in STEM fields.

The study builds upon and extends existing research in two ways. First, rather than look solely at the rates at which students enter, persist, and complete academic majors in STEM fields, this study examines students' actual pathways toward degree, taking into consideration the courses taken and differences in course taking patterns between female and male students who may be considering an academic major in a STEM field at a four-year university. Second, this study leverages new analytic methods to sequence course taking patterns – over students' trajectories in college – to better understand the dynamic processes underlying differences in persistence and completion rates in STEM-related fields.

CHAPTER 2: LITERATURE REVIEW

In the past few decades, due to the increasing importance of STEM fields for the Nation's economy, there has been significant investment in improving STEM education at different points in the educational pipeline. For example, the National Science Foundation has funded multiple projects to revolutionize engineering and computer science departments throughout the US (Chen, 2015). This focus and investment, consequentially, has attracted researchers from different fields and led to growth in research on STEM education. Much of the literature in higher education has been descriptive, documenting enrollment, retention, and attrition rates and examining association between individual and institutional characteristics and educational attainment in these fields. I mentioned these studies in the previous chapter to provide the reader with background information on how college students are doing in STEM fields. Recently, research in STEM fields has shifted toward finding non-demographic factors that contribute to persistence and disparities in STEM fields. Researchers have examined a range of contributing factors. In this section, first, I review general factors that contribute to college students withdraw from STEM fields. Then, I consider research that specifically examines factors that contribute to female students' under-representation in these fields. Finally, since my goal is to look at persistence and disparity in STEM through analyzing students' academic behavior, I look at the research which have taken different methodological approach to explain students' academic behavior.

Factors Contributing to Persistence in STEM Fields

Research investigating factors that influence students' decision to leave STEM fields may be broadly organized in the following categories: 1) academic preparation; 2) institutional factors; and 3) performance in "gate-keeping" STEM courses.

Academic preparation for college level coursework has been identified as one of the key predictors of student persistence toward a college degree in a STEM-related field. Numerous studies suggest that indicators of a student's preparation for college – such as taking Advanced Placement (AP) courses in STEM content areas in high school and having higher grade point averages and admissions test scores – are associated with persistence and degree attainment in STEM fields (Chen, 2013; Griffith, 2010; Kokkelenberg & Sinha, 2011; Tyson, Lee, Borman, & Hanson, 2007). For example, Kokkelenberg and Sinha (2011) find that taking more STEM AP classes in high school is associated with an increased chance of graduation with a STEM degree, and Chen (2013) finds that having a high school GPA of 3.5 or higher significantly decreases the chance of switching to a non-STEM field.

Research also suggests that the type of higher education institution a student attends may also influence their persistence toward a degree in a STEM-related field (Chen, 2013; Griffith, 2010). STEM entrants who first attended highly- or moderately-selective institution are more likely to pursue a degree in STEM fields than their peers who attended less selective institutions (Chen, 2013). Among selective institutions, those with a large graduate-to-undergraduate student ratio and that devote a significant amount of spending to research have lower rates of student persistence toward degrees in STEM fields (Griffith, 2010).

Campus environment may also affect student persistence in STEM fields, especially for non-White and female students (Chang, Sharkness, Hurtado, & Newman, 2014; Hurtado et al., 2007; Marx & Roman, 2002; Ost, 2010). For example, Chang et al. (2014) found that institutions that engage students in academic experiences such as studying frequently with others, participating in undergraduate research, and involving students in academic clubs or organizations increase underrepresented students' persistence in STEM fields. Moreover, Hurtado et al. (2007) found that perceptions of hostile racial climates negatively impact minority student adjustment and integration in STEM-related academic majors.

Research on factors determining persistence and graduation from college with a STEM-related degree points toward the number of STEM courses taken in the first year of study, as well as the type of introductory STEM courses taken (especially math courses) in that year, are closely linked to a successful completing a degree in STEM fields (Adelman, 2004; Chen, 2015; Crisp, Nora, & Taggart, 2009; George-Jackson, 2016; Ost, 2010; Rask, 2010; Seymour & Hewitt, 1997). Chen (2013) finds that students who persisted in STEM fields earned an average of 18 STEM credits in the first year of their study. She also found that a proportionally higher number of those students took calculus or other advanced mathematics courses in their first year of study compared to their peers who left the fields (81 percent vs. 36 % of STEM leavers who left college and 57 percent of STEM leavers who switched to non-STEM majors).

A student’s performance in entry-level STEM courses in his/her first year of study also influences the decision to stay in or leave STEM fields. Research shows that poor performance in STEM courses, especially relative to performance in non-STEM courses, leads to students’ switching to non-STEM degrees or leave the university entirely (Chen, 2013; Ost 2010; Rask, 2010). Chen (2013) found that a higher percentage of STEM leavers who dropped out of college or switched majors earned at least one grade point higher in non-STEM courses than STEM compared to their persistent peers. In another study examining high-performing students’ attrition rate in STEM fields, Chen (2015) finds that the probability of switching majors for high-performing students was associated with poor performance in STEM courses and she suggests that one of the motivating factors for students to switch to degrees in non-STEM might be due to their experiences in initial STEM courses.

While a range of factors contribute to students’ persistence toward degrees in STEM fields, some researchers argue that factors such as student performance in entry STEM courses play a more important role in students’ decisions than do other factors. Student performance in entry-level STEM courses that are intended to sort students into STEM and non-STEM degrees – i.e., “gatekeeper courses” – have been identified as a key indicator of whether a student will successfully graduate with a degree in a STEM field (Adelman, 2005; Chang, Cerna, Han, & Sáenz, 2008; Seymour & Hewitt, 1997). Research shows that controlling for performance in these courses weakens the effects of other factors on the likelihood that a student completes a degree in a STEM field (Ost, 2010; Rask, 2010). Consequently, some researchers recommended that future research in persistence needs to prioritize exploring students' STEM coursework in college, especially students' dynamic course-taking process (Chen, 2013 & 2015; Shapiro & Sax, 2011). To date, however, only a few studies (e.g., Wang, 2016) have considered student course taking patterns, particularly the sequence in which courses are taken, and the likelihood that students complete a degree in a STEM field.

Gender Disparity in STEM Fields and Contributing Factors

Despite the fact that during the 2014-15 academic year women made up more than half of college students (57percent) nationwide (National Science Foundation, 2016), females are significantly underrepresented in population of students who obtain a college degree in a STEM field, especially in engineering, computer, and physical sciences. Based on National Science Foundation report, women made up 18.4 percent of the undergraduate population in engineering (National Science Foundation, 2016). This underrepresentation is a challenge, with serious consequences for the society and the economy. Finding ways to increase women's participation in STEM fields could benefit the fields themselves and the overall economy by bringing more creativity and diversity of ideas to the workforce.

Many attempts have been made to understand factors that contribute to this underrepresentation, its persistence, and to find ways to attract more women to these professions (Davis et al., 1996; Fox, Sonnert, & Nikiforova, 2009; Griffith, 2010; Hill et al., 2010; Hughes, 2011; Seymour, 1995). Researchers have offered a number of theories to explain this disparity (Ceci & Williams 2010; Hyde et al., 2008; Lynn & Irwing, 2004; Seymour, 1995; Tyson et al., 2007). These include theories that are based on biological difference (e.g., Lynn & Irwing, 2004), academic preparation (e.g., Chen, 2013), negative attitude (e.g., Weinburgh, 1995), absence of role model (e.g., Hill et al., 2010), STEM curriculum, pedagogy (Davis et al., 1996), and cultural and social stereotypes (e.g., Steele, James, & Barnett, 2002).

For a long time, a deterministic framework, based on the premise of women's intrinsic inability in math and science, was used to explain women's absence in STEM fields (Ceci & Williams, 2010; Lynn & Irwing, 2004). This view, however, has been challenged by research that shows comparable aptitude between female and male students (Burger et al., 2007; Hyde et al., 2008; Tyson et al., 2007). For example, differences in math and science competency within each gender are far larger than the average difference between the sexes, and other studies have found very little difference in scientific or mathematical ability between the sexes (Blickenstaff, 2005).

As mentioned earlier, academic preparation in high school is an important predictor of STEM persistence in college. Thus, there has been speculation that women's underrepresentation in these fields, especially in math and engineering, might be a result of their differential preparation in mathematics and sciences in high school. Various studies, however, reject this hypothesis, pointing to the fact that girls earn math and science credentials at the same rate as boys do, and frequently even earn better grades in high school coursework in these subjects (Tyson et al., 2007; Voyer & Voyer, 2014). This suggests that women are at least equally, or perhaps even better prepared, to pursue a STEM major in college.

Taken together, existing research suggests that women's low enrollment in and attrition from college degrees in STEM fields cannot be explained by the measure of their ability and preparation. This suggests other motivational, social, and institutional factors likely explain under-representation of women in the share of students attaining a college degree in a STEM field. For instance, some studies have found that girls have a negative attitude toward math and science compared to their male peers, and that these negative attitudes contribute to their decision not to pursue a college degree in a STEM field (Riegle-Crumb, Moore, & Ramos-Wada, 2011; Weinburgh, 1995). Others, however, argue that these disparities in attitude and motivation toward science and math could not be considered independent factors (Blickenstaff, 2005; Burger et al., 2007; Hill et al., 2010). Instead, attitude and motivation are the result of sex-role socialization and closely tied to other social and environmental factors that make such subjects unattractive to girls (Pinel, Warner, & Chua, 2005).

Drawing on a large body of research, Hill and colleagues (2010) provide evidence that negative stereotypes about women's ability in math and science persist and that they significantly impact women's attitudes, self-assessment, and aspirations in pursuing a career in STEM fields. Felder, Felder, Mauney, Hamrin, and Dietz (1995) found that female students were more likely than men to attribute their poor performance in STEM courses to their own lack of ability, while men were more likely to attribute it to a lack of hard work or being treated unfairly. A survey from freshman female college students found that, in spite of their academic advantages, females rated their academic ability and creativity lower than their male counterparts (Almanac, 2016). There also is evidence that women’s low self-assessment, aspiration, and motivation are in part caused by negative interactions they have with their peers and professors. For example, studies have found that undergraduate females in STEM courses feel that their faculty and male peers do not take them seriously (Neumann, Lathem, & Fitzgerald, 2016; Shapiro & Sax, 2009; Sprecher, Brooks, & Avogo, 2013). Such negative interactions are themselves the result of implicit bias in associating strength in math and science fields with being male. This, in turn, impacts women’s academic aspirations and performance, and consequently their persistence toward a degree in STEM fields.

In addition to these negative stereotypes and biases, there are institutional barriers that act as a gender filter that obstruct women’s path in pursuing a major or career in STEM (Fox et al., 2009). Women can face a chilling climate in postsecondary classrooms, ranging from outright hostility, harassment, and verbal abuse, to calling on and encouraging men more often than women (Burger, 2007; Hill et al., 2010). This chilling climate lowers even highly-skilled and motivated women’s sense of belonging to the academic environment, which leads to their isolation and a feeling intimidation among other feelings (Walton et al., 2015). Working in such an unwelcoming climate may put women at a higher risk of switching to other fields or even dropping out of college (Shapiro & Sax, 2011; Walton, Logel, Peach, Spencer, & Zanna, 2015).

Pedagogy and curriculum are two other institutional factors that have been associated with women deciding not to pursue a degree in a STEM field (Blickenstaff, 2005; Shapiro & Sax, 2011). Research shows that a competitive and aggressive nature of pedagogy in STEM courses that emphasizes individual success rather than collaborative learning may discourage women from taking courses or pursuing a degree in STEM fields (Seymour, 1995). Women also report finding the curriculum in STEM fields impersonal and irrelevant to human condition, which negatively impacts their academic aspirations (Beyer, 2014; Burger, 2007). In addition, introductory courses' failure to provide a holistic view to subject area instead portraying science and engineering as highly competitive and masculine domains also may filter women in the curricular process and redirect them to non-STEM fields (Blickenstaff, 2005; Fox et al., 2009).

Finally, the lack of role models for women in STEM fields can discourage women from pursuing a degree in these fields. Unfortunately, as Shapiro and Sax (2011) explain, due to the fact that female faculty are underrepresented in STEM department, female students have limited access to same-sex role models and mentors. This may discourage women from pursuing a career in these fields or send a message that women do not belong to these fields. In their recent study, Neumann et al. (2016) found that having women role models played an important role in women's persistence in engineering departments. Female role models helped women see what being successful looked like for a woman like them in engineering.

In summary, researchers have been able to identify a range of factors that contribute to women's selection, pursuit, and attainment of STEM field degrees. In particular, institutional (Fox et al., 2009) and structural barriers play a role in women losing interest in STEM majors, especially engineering, computer science, mathematics, and physical sciences. These factors are in addition to ones that influence all students' decision, irrespective of their gender, to stay or leave such fields. Together, these factors help explain why women comprise a smaller share of students who enroll in and graduate with degrees in STEM fields.

Student Academic Behavior

One of the characteristics of postsecondary education in the US is the diversity of pathways students could follow through their study. Most students entering college do not declare their major until the third or the fourth semester (Shapiro & Sax, 2011). During the first semesters they take different courses offered by programs and try to find their way for declaring a major. Even after declaring a major, it is not unusual for students to decide to switch to other majors. We know that a student's decision to declare a major, stay in one, or leave it is influenced by choices made at different points of his/her college career, under different circumstances. For example, encountering difficult or disengaging courses or getting poor grades in particular courses might cause some students to redirect their efforts to another major or sometimes even cause them to drop out of the college or transfer to another institution (Adelman, 2006; Chen, 2013; Seymore, 2002). That is to say, a student's curricular experience is a dynamic process that influences her/his subsequent course-taking decisions as well as the progress toward selecting, or changing, his/her major, and finally, the completion of degree requirements.

The college curriculum in any given major is an academic plan developed and structured by faculty, program directors, and the administration with the goal of enhancing students learning and achieving a certain level of literacy in a given field. The experience of interaction with the curriculum is a complex and multilayered one influenced by different components of the curriculum including content, pedagogy, and instructional resources, the faculty, and other external factors (Cohen & Kisker, 2012). Therefore, understanding students' curricular experience is crucial for evaluating how successful the institution has been in fostering students' learning. Although numerous studies have corroborated the importance of examining student academic behavior (Adelman, 2005; Chen, 2015; Shapiro & Sax, 2013), few efforts have been made to investigate the dynamic feature of experience over time. In existing research, the focus has been on pathway analysis, which is based on college students' persistence framework. In such a framework, pathways are conceptualized as outcomes and measured with dichotomous (complete/disrupted) variables. They also identify a set of proximal variables in their model in the hope of explaining student academic behavior throughout the college.

In recent years, researchers have started applying more advanced methods in which pathways are measured in categorical (complete/ part-time/ discontinuous), rather than dichotomous variables. They have also included independent variables with several data points between college entry and exit to better show students' academic behavior (Chen, 2013, 2015; Ewert, 2010). Although this framework provides valuable information, as the researchers themselves acknowledge, these studies have serious limitations in capturing the full picture of student behavior. Even the most advanced of these methods are not able to reveal the complex interaction between course taking experiences across time due to the assumption that is at their foundation, i.e., the linearity and uniformity of student behavior (Bahr, 2013). As a result, student pathways towards earning a degree and the influence of different pathways on their outcome has remained understudied. We are left with little knowledge about whether taking different pathways align with successful outcomes or whether different group of students, such as women, who are underrepresented in certain fields such as STEM, are inadvertently led by the system to take different pathways that translate to a different degree of success.

An additional area of problem with studies investigating STEM pathways is that they identify STEM students based on their major reported in the beginning of their study. However, students often do not declare their major until their junior year. Therefore, such studies overlook a considerable number of students who might have intended to major in STEM but after difficult experiences with initial STEM courses decided to redirect their studies to another major. In persistence studies, these students are of great interest and excluding them could lead to misunderstandings. To reach a complete and more accurate picture of the STEM pipeline instead of just tracking students by their declared major field, researchers could use transcript data, which provides a road map of majors and courses taken by each student throughout her/his study in college (Shapiro & Sax, 2011).

There have been a few research efforts that apply new analytical approaches, particularly data mining, to examine questions related to academic behavior and try to find ways to identify different pathways students take to go through their academic programs. Although not all are conducted at the college level or are related to STEM fields, they are relevant to this study because of their common goal – i.e., to identify students' academic behavior using detailed student transcript data. I will examine these studies below. I will also consider other studies that rely on simple descriptive analysis.

The reason I mention them here is due to their innovative approach in using detailed data and their influence on course-pattern identification studies. For example, most studies conducted by Adelman (1999, 2004, 2006), were very influential in highlighting the power of transcript-base analysis. His focus on student academic history inspired new lines of research both in community college and four-year college context by others. Therefore, it is essential to include them in this review.

In this section, I will review these studies in more detail, laying out what has been done in this field and, more importantly, what is needed to be done to identify students' course-taking patterns and significant differences in course-taking patterns by different student groups.

Student course taking behavior. Friedkin and Thomas (1997) completed one of the earliest studies of student course taking behavior and were among the earliest to propose the idea that differences in student educational attainment accumulate over time and may be understood as arising from differentiated patterns of coursework taken in a multiyear sequence of schooling. In their study, the authors develop a theoretical rationale for viewing course-taking patterns as student social positions in students' relations with particular teacher and coursework during their high school years. They then applied this framework to analyzing a nationally-representative sample of high school students who then proceed to college using data from U.S. Department of Education's High School & Beyond Survey. Employing network analysis to the profiles of high school students' coursework, they find distinct profiles that conform to most students' course profiles. Then, they use hierarchical cluster analysis, which results in eight curricular positions of students' course-taking patterns. Each student is then assigned to the closest matching curricular position. After this, Friedkin and Thomas (1997) investigate the association between student characteristics with the curricular position they have been assigned to, finding out that students' unique membership to any of the eight positions were associated with their demographic status, academic skills, and achievement. They conclude that even without a formal system of tracking, by the end of their schooling the students would be differentiated with respect to their course-taking patterns.

Heck, Price, and Thomas (2006) extend this study by applying the same analytical approach to a set of transcript data from a comprehensive high school instead of using surveys from samples of students. Seven distinct course-taking patterns with a high degree of fit emerges from their network analysis. Further analysis of the characteristics of the student members of the profile suggest that students are dramatically differentiated by the seven profiles based on their socioeconomic status, ethnic/racial groups, and their academic outcomes, demonstrating wide inequalities in students' outcomes and aspirations.

The Friedkin et al. (1997) and Heck et al. (2006) studies are exemplary as they present a new approach and emphasize the importance of understanding differences in students' educational attainment from the perspective of differentiated patterns of coursework in multiyear sequences. Although both studies began with high ambitions to include all dimensions of the data in their analysis, in later stages of research they dropped the element of time/semester in which a particular course had been taken and also the teacher(s) who taught the course due to the difficulty of handling a large multidimensional dataset. To date, however, the conceptual frameworks and analytic approaches developed in these studies have not been applied to higher education.

In another effort, Adelman (1999, 2004, 2005, 2006) has conducted various studies in which he uses longitudinal students' transcript level data to illuminate paths to degree completion in two/four-year colleges. Although his studies mostly focus on simple descriptive analysis, they have been very influential in highlighting the power of transcript data in understanding students' academic behavior and determinant factors contributing to their success. For example, in *Answering the Tool Box* (1999) and *The Tool Box Revisited* (2006), his transcript-based analysis reveals the determinant role of early momentums, such as taking a number of college-level math courses as early as possible, on degree completion. In another study, *Moving into Town* (2005), Adelman uses transcript data to classify traditional-age community college students based on their academic history and the number of credits they earn from community college. Although most of these studies were descriptive, they were influential in inspiring several paths of research based on transcript analysis, especially in community college settings.

Community college's diverse student population, their institutional flexibility to choose how and when to enroll, and their path of studies and transferring to college have made it an appealing case to apply pathway analysis. A number of studies in the recent years have conducted in community college to identify students' pathways. Most of these studies aim at identifying the typology of its students. Influenced by Adelman's uses of student transcript data, they began applying new methods to transcript data to develop student typology. For example, in *The Bird's Eye View of Community College*, Bahr (2010) develops a behavior typology based on students' course-taking patterns and other enrollment patterns using K-mean cluster analysis. He identifies six clusters of behavior including: transfer, vocational, drop-in, noncredit, experimental, and exploratory. Then, by examining the relationship between students' demographic characteristics and cluster membership, he explores whether different group of students have different course-taking behavior. In another study, Zeidenberg and Scott (2011) use transcript data from Washington State community college system to investigate students' course-taking patterns. They apply Partitioning Around Medoids (PAM) clustering separately to liberal arts and career-technical (CTE) students to organize students to groups based on similarity of courses they have taken. Their cluster analysis results in 20 solutions of course-taking patterns in CTE subsample and 5 solutions in liberal arts. Then, to discover what type of students are in each program they examined the demographics and the completion and transfer rates of the students within each cluster. The authors conclude that clustering would be useful to researchers throughout education who are trying to understand student course-taking patterns using a large-scale transcript data. They also acknowledge the limitation of their study in analyzing course-taking activity without considering their sequential order and plan, in the future, to look in more detail at the sequencing of this taking-course activity.

In another study, Bahr (2013) criticizes the traditional dominant input-output analysis approaches in community college students' research, which is heavily focused on examining the relationship between community college students' characteristics and their outcome. He argues for the necessity of developing a new approach to capture various pathways and behaviors. After showing the limited capacity of traditional input/output analysis in providing information on how and why some college students fail or progress through the college, he presents a new deconstructive approach to illuminate community college students' pathways and the relationship between these pathways and student outcomes. Bahr (2013) argues that his new framework deconstructs "the varied steps or stages through which students pass from the point of the college entry to a given outcome of interest…In other words, this approach constitutes a shift from the focus on outcome that has dominated research on community college students to focus on process" (p. 145).

Influenced by Bahr and Aldeman's studies, and in what can be considered a major step forward in the last couple of years, few researchers have started applying more advanced analytical techniques to identify patterns that align with degree earning or transfer in a community college setting. For example, in *Intensity and Attachment: How the Chaotic Enrollment Patterns of Community College Students Relate to Educational Outcomes*, Crosta (2014) tries to identify community college students' behavior patterns using students-level transcript data from several community college campuses. His study is similar to other studies in community college that aim at identifying the typology of its students. What makes Crosta's (2014) study different, however, is his focus on clustering longitudinal patterns created by intensity and continuity of students' enrollment instead of a set of variables. To identify the patterns, he creates an enrollment vector for each student that consists of zeros, ones, and periods for 18 semesters. Using visualization techniques of enrollment patterns, which is very unique to this study, he visualizes the entire range of enrollment. Then, using a K-mean clustering technique, the author identifies six clusters of enrollment patterns. The clusters emerge only from students' sequential enrollment patterns without using any other information. The results from cluster analysis identify Early Leavers as the most common pattern among community college followed by Full-time Persisters and Early Persistent Switchers. Crosta's (2014) study is unique in using longitudinal patterns and visualizing those patters in a way that really helps to better understand common student enrollment patterns. His study, however, does not provide any information about course-taking patterns. That is, the question of whether taking specific courses lead to a student's decision to leave his/her studies early or stay in the college remains unattended.

In another study, Wang (2016) uses Bahr's deconstructive framework to explore course-taking patterns of community college students. Her study is one of the few that uses various advanced data mining techniques for exploring students' course-taking patterns. Although other studies use primary data mining techniques such as clustering, her study is unique in the fact that it is the first to argue for the necessity of using data mining techniques, justifying their application to transcript data. She also uses various techniques to provide a comprehensive analysis of community college students' course-taking patterns. In this study, Wang (2016) makes a strong argument that due to the complex and unstructured nature of transcript data, which consist of tens of courses recorded for each student over a number of academic semesters, data mining techniques can provide better insights into a student's academic behavior.

The purpose of Wang's (2016) study is to identify course-taking patterns that have been successful in transferring community college students to four-year colleges in STEM fields. To achieve this goal, she applies *frequent pattern/ association rule mining* technique to the Beginning Postsecondary Students Longitudinal Study (BPS: 2009) to identify frequent course-taking patterns. Every student's course-taking pattern constitutes various itemsets. Each itemset is a set of courses taken by a given student in one semester. Using Apriori Algorithm, she identifies the frequent course-taking patterns that result in three different outcomes: transfer to STEM, transfer to non-STEM, and non-transfer. Then, she applies Decision List Algorithm to add other predictors variables, such as the dosage of particular courses that have been taken by a student. Finally, to add student demographic characteristics to the analysis, Wang (2016) applies Decision Tree algorithm to examine the relationship between those characteristics and course-taking patterns.

Wang's (2016) pattern mining results provide unique insights into community college students' trajectories to STEM transfer pipeline that would have not been uncovered by any of the traditional analysis methods. For example, one of the most striking results that emerge from her examination of these patterns is that in math-learning paths, math course-taking during the first semesters does not appear as a frequent pattern among transfers to STEM paths. Instead, taking "likely transferable" courses during the first semesters, followed by math courses in the subsequent semesters, is the most viable path to transfer to STEM. In fact, the math-learning path is the most common feature of transfer to non-STEM patterns. Such important information on the patterns that contribute to a successful transfer to STEM could be used by program designers and advisors in community colleges to improve and facilitate student outcome. The most valuable contribution of the study, however, is highlighting the importance of utilizing data mining techniques to analyze rich transcript data that is available to broaden our understanding of students' academic behavior.

An important study that presents an innovative approach in utilizing data mining techniques to student's map of study, Wang's (2016) study has its shortcomings. Just like most of the previous studies, Wang (2016) fails to take into account the sequential feature of a student's course-taking pattern. A student's course-taking pattern is a sequential pattern, meaning that it is an ordered list of sets of courses taken by him/her over the time of study. Ignoring this important feature restricts, and might even distort, our analysis of students' academic behavior.

Another study that applies new data mining techniques to longitudinal transcript data (Beginning Postsecondary Students Longitudinal Study: 2004/2009) to examine course taking patterns' contribution to degree completion at college level is Witteveen and Attewell's (2016). They apply Hidden Markov Model (HMM) to transcript data collected from a sample of U.S. four-year college students in order to predict degree completion and non-completion. Their goal is to build a model that effectively recognizes a graduating or non-graduating student after only one or two years of college transcript information. HMM is a new data mining technique used to identify the hidden states that are associated with both static observable states and hard-to-observe trajectories leading to particular outcome states.

Their initial analysis for building the HMM model suggests a combination of six to eight variables associated with a “three-state solution” as the most effective model for creating a coherent and distinct state description. Initial states for graduating students include: state 1: high credits, state 2: high STEM, state 3: STEM/withdrawing. For non-graduating students the three states are: state A: low activity, state B: low STEM, and state C: STEM/high credits. The authors then analyze the probability of moving to future states given the knowledge of any current state. With regards to graduating students, the results suggest that they rarely take STEM courses in combination with a large number of credits. Rather, they withdraw, or they take fewer courses when attending technical courses. Their model, however, has difficulties in distinguishing non-graduating students and does not offer that much insight into their trajectories.

Witteveen and Attewell’s (2017) analysis of the association between socio-economic factors and college states indicates the consistent and significant effect of *gender*, predicting that male students are more likely to be in a “STEM/high credits” state. In contrast, other demographic and high school variables are not significant predictors of HMM states. Their study’s results offer valuable insights into the complex interaction between course-taking experience over time, which again could not be captured by traditional linear modeling. Their study also corroborates the need for utilizing more advanced data mining techniques such as HMM when detailed transcript data is available.

As this literature review reveals, understanding student course-taking patterns and its influence on outcome has attracted more attention in recent years. Many researchers

(Adelman, 2005; Bahr, 2013; Chen, 2015; Shapiro & Sax, 2013; Zeidenberg & Scott, 2011) have issued calls for the use of new analytical approaches to find answers for various questions related to college students' academic trajectories and their influence on subsequent outcomes. Such calls, however, have been answered only by few people and, as a result, the move toward bringing new approaches to explore these areas of research has been slow. In other words, most research in the field is still conducted using traditional approaches. To tackle this issue, it is my intention in this study to propose a new data mining technique applicable to students' transcript data in order to identify and understand their course taking patterns in STEM fields in a four-year college setting. The few studies that aim at a similar goal, that is, identifying student academic trajectories and course taking pattern, have mostly focused on community college setting. Four-year college students' academic paths and course-taking patterns, especially in STEM fields, have rarely been touched by scholars. Identifying these patterns can help us not only to understand paths that lead to STEM major selection and ultimate degree attainment but also to identify particular type of courses or sequence of courses that may act as gatekeeper, leading some students to leave their field by switching to other fields or dropping out of the college.

As mentioned above, scholars have debated whether there is a gender disparity when it comes to persistence in STEM fields. It has been argued that some of the disagreement on the topic might be the result of failing to consider the persistent pattern differences between soft and hard sciences (Kokkelenberg & Sinha, 2011; Ost, 2011; George-Jackson, 2016). My research approach could offer important insights for us to settle this question. Using data mining techniques, I can find out whether female students who decide to pursue a major in STEM fields take significantly different paths compared to their male peers and, if so, at what point of their study, or after taking which sequences of courses, this departure begins.

CHAPTER 3: DATA & METHODS

The purpose of this study is to examine academic pathways through college among students who may be considering a STEM-related major. First, I describe the share of students with declared majors when they first matriculated as a degree seeking student at the university. I then explore how this initial distribution of students' academic majors changes over time, and the extent to which patterns differ for female and male students. To do so, I employ two different data mining techniques – Sequential Pattern Mining and cluster analysis of academic major sequences. Both techniques provide a somewhat different perspective on students' academic experiences, as well as a useful comparison among potential methods for exploring patterns in higher education students' academic major trajectories. Specifically, I consider three research questions:

1) What major and course-taking patterns are aligned with successful degree attainment in STEM fields?
2) Are there any significant differences between men and women's academic major and course-taking patterns within STEM fields?
3) In which phase of their program of study, do STEM students switch to other fields?

To answer these questions, I leverage recent developments of data mining techniques and apply these strategies to coded undergraduate transcript data at a four-year university. My approach in identifying student course-taking patterns is similar to techniques used in market basket analysis to identify costumer shopping behavior. Using Sequence Pattern Mining techniques, I create sequences of academic majors and STEM related courses a student takes each semester. Then, by clustering those sequences I identify patterns of academic major and course-taking that are common to students who successfully pursue the fields and obtain a degree. Such patterns could reveal which groups of students have taken similar academic major paths and also which groups of students have decided to switch to other fields or dropped out of college after declaring a major in STEM fields. In terms of students' course-taking patterns, they will reveal the sequences of STEM courses, and the characteristics of student group that stopped taking STEM courses. As a result, I can identify the so-called "gate-keeper courses" that compel certain groups of students not to take any further STEM courses, switching to another field or dropping out of the college. I believe identifying course-taking patterns has a great potential policy implication for designing and structuring effective pathways and developing efficient interventions to improve student retention in STEM fields. Therefore, the study's findings hold potential to influence decision making by a broad group of stakeholders in higher education, including students, educators, and administration.

Data

Full population. Data were provided by the University's Office of Institutional Research and included student transcript information for three cohorts of students of students for a period of six years after their initial matriculation to the University (2010, 2011, and 2012). In total, there were 9,086 students, across the three cohorts. Table 3.1, Column 1 describes the demographic characteristics of these students. When we compare the makeup of the student body under examination in this study to national averages, a few facts stand out. First, female students comprise 56 percent of the student body in this study, a percentage consistent with national trends (Alamance, 2016). Second, in terms of racial diversity the makeup the student population considered in this study diverges from national averages – i.e., students included in this study were predominantly white (87 percent). Third, while nationally around 80 percent of students attending public universities have state residency, only 31 percent of students in this University were in-state residents. Finally, 75 percent of students completed their degree within six years of matriculation to the university, whereas, nationally, about 60 percent of students attending four-year institutions complete a degree within six years (NCES, 2016).

The data provided by OIR contained detailed transcript records, including a record of each course attempted by a student while enrolled at the University (during the six year time period considered for the study), for each of concurrent 12 semesters. In addition, the transcript data identified a student's declared academic major in each semester and the grade obtained for each course in which a student enrolled. Altogether, the dataset contained 415,200 course records for three cohorts of students who attended the University between 2010 and 2018.

Table 3.1

Descriptive Statistics for 3 Cohorts Entering the College 2010-2012

Demographic Characteristic	**All Students (Column 1)**	**STEM-Considering (Column 2)**
Female	56%	54%
White	84%	84%
Black	1%	1%
Hispanic	4%	3.6%
Asian	2%	2.5%
American Indian	0%	0%
Two or More Races	2.5%	2.7%
Nonresident Aliens	2.1%	2%
Unknown	4.3%	3.7%
State Resident	31%	31%
Transfer	20%	15%
Completed Degree	75%	78%
Total	9,086	4,890

STEM-considering population. There are different definitions of what constitutes a STEM field. For example, The National Science Foundation (NSF) has a broad definition that even includes social sciences. In this study, however, I use a narrower definition suggested by the National Center for Educational Statistics (NCES) that classifies the following fields as STEM: mathematics, physical sciences, biological/life sciences, computer and information sciences, engineering and engineering technologies, and science technologies. A detailed list of majors and course subjects that can be classified as STEM majors and courses based on this definition is provided in Appendix A. I identified non-STEM courses based on the subject of course provided in the data and recoded all of those courses to a non-STEM binary variable. Given the focus of this study, I was most interested in students who either declared a major in STEM upon matriculation to the University and also those who did not initially declare a major but who demonstrated an initial interest in STEM as evidenced by course taking patterns. While it is easy to identify students who declared a STEM major upon entrance, I needed to develop a criterion to determine whether a student was considering such declaration or switching to STEM. While it is impossible to develop a perfect criterion, it is reasonable to assume that a student who considered a STEM major or might switch to one would take STEM courses as they weigh their decision. Therefore, I also included in my analysis "STEM Considering" students, who took more than two STEM-related courses in their first year but who had not initially declared a STEM major. Using this logic, I created a second analytic sample, consisting of 4,890 students (2,625 female and 2,265 male). I refer to this group of students as "STEM Considering".

Student transcript data shows that there were around 750 STEM courses taken by students, in 50 subjects. Some of these courses were general introductory STEM courses, which a large number of students from different STEM programs took them. These courses were of special interest to me since there is discussion in the literature about some students leaving the relevant fields after taking them. I included all these courses in my analysis by their unique course subject and number. Other courses were only taken by students who majored in a specific field. I classified these courses based on their subjects and then, depending on whether they were introductory or advanced level courses, I assigned them as "Int" or "Adv." The final list of course categories included in the analysis is provided in Appendix C.

Analytical Approach

Using Bahr's (2013) deconstructive approach that calls for an in-depth analysis of transcript data to illuminate student academic trajectories and the relationship between these varied trajectories and student persistence, I propose a new method to identify various academic trajectories that lead to completion of a STEM major, switching or leaving the college. We know that a student's decision to declare a major, stay in one, or leave it is influenced by choices made at different point of his/her college career under different circumstances. One important factor is how the student interacts with the curriculum and his/her experience of such interactions. Detailed student transcripts are an important piece of multidimensional data, which can provide us with valuable insights into the student's experience in navigating the curriculum and interacting with it and how it influences her/his decision-making process in following different academic paths overtime. Some of the pathways lead to progress through the college years and the eventual completion of the program of study. Other pathways lead to failure. We know, however, very little about the various academic trajectories that students take to go through their study and how they influence a student's outcome. Identifying these paths could offer a lot of information about courses or sequence of courses that enable students to successfully take a path to choose a major or earn a degree. They can also reveal courses that play a gatekeeping role in preventing some group of students from going further in their program of study and their decision to switch to a different field or leave the college.

My analytic approach took into account the multidimensionality of student transcript data. To identify academic trajectories, former studies have focused on students' academic majors at the point of entry or the accumulation of courses during the program of study. They did not consider the longitudinal sequence of students' academic experience. Using Sequential Pattern Mining, I identify the most frequent patterns for academic majors over students' enrollment periods. This helps us understand how and when students change academic majors. Applying this method to students' course profiles, I identify frequent course-taking patterns and patterns that are aligned with degree completion. Also, since women have been underrepresented in STEM fields, especially in fields such as engineering and computer science, I examine whether female and male students follow different academic major and course-taking patterns within STEM fields. Using Sequential Pattern Mining techniques, I also identify sequence of courses that increase the probability of leaving the program to a non-STEM or dropping out of the college.

Methodology

In the past few decades, and with the emergence of fast-growing technologies that have made collecting, storing, and processing large amount of data possible, multidimensional data have become available at a large scale for researchers in various disciplines such as bioinformatics, finance, geology, and marketing (Dong & Pie, 2008). In this context, new methods and techniques have emerged that enable analysts to unpack the complicated structure of data and discover, or "mine", hidden knowledge in large datasets (Kantardzic, 2011). In response to such demands, fields such as data mining have rapidly developed, offering researchers new techniques to effectively manage and analyze such data (Dong & Pie, 2008).

Data mining, also known as Knowledge Discovery in Database (KDD), is an analytical process of discovering consistent and useful patterns and relationships hidden in a large-scale dataset (Dong & Pie, 2008). Unlike traditional hypothesis testing designed to verify *a priori* hypotheses about relationships between variables, data mining is used to identify systematic relations between variables when there are no, or incomplete, *a priori* expectations as to the nature of those relationships. Data mining has the advantage of imposing very little in the way of prior assumptions about what is in the data; rather, it allows the data to tell the researcher what is going on (Han et al., 2011). In a typical data mining process, many variables are accounted for and compared, using a variety of techniques in the search for systematic useful patterns (Han, Pei, & Kamber, 2011).

When it comes to the primary goal of data mining tasks, data mining constitutes a range of techniques from descriptive, on one hand, to predictive on the other. On the predictive end of the spectrum, the goal is to produce a model that can be used to predict unknown or future values of variables of interest. *Classification*, *Regression*, and *Dependency Modeling* are examples of predictive data mining tasks (Kantardzic, 2011). *Descriptive* data mining is focused on finding useful interpretable patterns and relationships that describes the data. For example, *Clustering*, *Summarization*, and *Change and Deviation Detection* are examples of descriptive data mining tasks (Kantardzic, 2011).

However, data mining techniques are still underutilized in educational research (Wang, 2016). Although we now have a new field, called educational data mining, with its own association and biannual conferences, most of existing research has been focused on E-learning and rarely deals with data from traditional educational settings. In my literature review, I found only a few studies that have used data mining techniques for research conducted in traditional educational settings (e.g., Witteveen & Attewell, 2017)

In this study, I used data mining techniques to identify course taking patterns from students' transcript data. Since student transcript data is a sequential, meaning that a student took different courses in semester order, the analysis needs to consider this sequential ordering of the data. Using pattern mining techniques like frequent item mining does not account for sequential data structures, and may fail to discover important patterns in the data or find patterns that may not be useful because they ignore the sequential relationship between semesters (Fournier-Viger et al., 2017).

Sequences are one of the important types of data that can be found in many domains such as medicine, biology, business, and other fields. For example, sequences are used to represent data such as sentences in texts (sequences of words), sequences of items purchased by customers in retail stores, and sequences of Web pages visited by users (Dong & Pei, 2007).

Sequential pattern mining is a data mining technique used to identify patterns of ordered events within a database (Han et al., 2011). First introduced in 1995 by Rakesh Agrawal of IBM's Almaden Research Center, its original application was in market analysis where it was used to predict whether within a certain time period after purchasing a certain product a customer is likely to purchase its sequel (Agrawal & Srikant, 1995). Soon, sequential mining techniques were used in different fields such as medicine, genetics, and marketing (Mooney & Roddik, 2013). That said, data mining techniques in general, and sequential pattern mining in particular, have not been widely used in educational research.

Sequence concepts. The order among the elements of a sequence may be defined by time as in event histories, or by physical positioning as in biological sequences or text sequences (Dong & Pei, 2007). Assume that I = $\{i_1, i_2, i_3, \ldots, i_n\}$ is a set of items. An itemset X is a set of items such that X ⊆ I. The notation |X| denote the number of items in an itemset X. An itemset X is said to be of length k or a k-itemset if it contains k items (|X| = k). A sequence is an ordered list of itemset s=$\langle I_1, I_2, \ldots, I_n \rangle$ such that $I_k \subseteq I$ $(1 \leq k \leq n)$. for example, itemset s_1={*Math I, Physic I, First-Year Seminar , Diversity*} contains four items, which are courses taken by a student in his/her first semester. The sequence $\{s_1, s_2, s_3, s_4, s_5, s_6, s_8\}$ represents the student's course-taking sequence (profile) for eight semesters.

A sequence $s_a = \langle A_1, A_2, \ldots, A_n \rangle$ is said to be a subset of sequence $s_b = \langle B_1, B_2, \ldots, B_m \rangle$ if and only if there exist integers $1 \leq i_1 < i_2 < \ldots < i_n \leq m$ such that $A_1 \subseteq B_{i1}$, $A_2 \subseteq B_{i2}, \ldots, A_n \subseteq B_{in}$ (denoted as $s_a \sqsubseteq s_b$). A given input-sequence database has the following fields: sequence-id, event-time, and the items present in the event. It is assumed that no sequence has more than one event with the same time-stamp, so the time-stamp may be used as the event identifier. The support of a sequence s_a, denoted as $sup(s_a)$, in a sequence database is defined as the number (or proportion) of input-sequences in the database that contain s_a (Dong & Pei, 2007).

Based on the type of items in a sequence, it can be categorized either as a state or event sequence (Ritschard, Gabadinho, Studer, & Müller, 2009). Here, a state, like full time residency status, refers to an item that lasts for a specific duration of time, whereas an event – e.g., taking a course – refers to an item that happens at a given point of time and has no duration. State sequences are useful for studying durations while event sequences are used for analyzing the order in which events occur (Ritschard et al., 2009). For instance, consider a student's sequential enrollment profile. If this is a state sequence, items could include student's major, residency status, or enrollment status (full/part time). If this is an event sequence, however, items can comprise of the courses a student has taken in a specific semester. An important difference between events and states is that multiple events can occur at the same time while states are mutually exclusive (Ritschard et al., 2009). For example, multiple courses could be taken by a student in a semester while he/she can't have both in-state and out-of-state residency.

Sequential pattern mining. Sequential pattern mining is the *task of finding all frequent subsequences in a sequence database that are common to several sequences* (Slimani & Lazzez, 2013). Those subsequences are called frequent sequential patterns. A sequence *s* is said to be a frequent sequence or a sequential pattern if and only if $sup(s) \geq minsup$. A minimum support threshold, set by the researcher, is a parameter indicating the minimum number of sequences in which a pattern must appear to be considered frequent and, thus, to be considered in the search (Founier-Vinger et al., 2016).

Numerous algorithms have been designed to discover sequential patterns in sequence databases. Some of the most popular ones are GSP (Generalized Sequential Patterns), Spade (Sequential Pattern Discovery using Equivalence classes), and PrefixSpan (Prefix-projected Sequential pattern mining) (Zhao & Bhowmick, 2003). All these sequential pattern mining algorithms take as input a sequence database and a minimum support threshold (chosen by the user) and output the set of frequent sequential patterns. In general, sequential pattern mining algorithms can be categorized as being either depth-first search or breadth-first search algorithms (Founier-Vinger et al., 2016). Breadth-first search algorithms such as GSP has been developed around this general idea that, if *s* is not a sequential pattern, we do not search any super-sequence of *s*, which is called Apriori property (Dong & Pei, 2007). A typical breadth-first sequential pattern mining method, mines sequential patterns by adopting a candidate subsequence generation-and-test approach based on the Apriori property (Dong & Pei, 2007). Given the database *S* and the minimum support threshold *minsupport*, the software first scans *S*, collects the support for each item, and finds the set of frequent items, that is, frequent length-1 subsequences. Then the frequent length-1 subsequence sets are used to generate new potential length-2 sequential patterns, called candidate sequences. Then, the sequence database is scanned again, and the supports of length-2 subsequences are counted. Those sequences passing the minimum support threshold are the length-2 sequential patterns. Using the length-2 sequential patterns, the set of length-3 candidates are generated. In the k-th pass, a sequence is a candidate only if each of its lengths -(k − 1) subsequences is a sequential pattern found at the (k − 1)-th pass. A new scan of the database collects the support for each candidate sequence and finds the new set of sequential patterns. The algorithm terminates when no sequential pattern is found in a pass, or when no candidate sequence is generated. The number of scans is at least the maximum i-length of sequential patterns. It needs one more scan if the sequential patterns obtained in the last scan lead to the generation of new candidates (Dong & Pei, 2007).

The challenge with breadth-first algorithms is their use of a very large search space to generate a huge number of candidate sets and constantly scan the database to discover the candidates (Slimani & Lazzez, 2013). To address this problem, depth-first algorithms – such as Spade, PrefixSpan, and FreeSpan – have been developed. Depth-first algorithms explore the search space of patterns by following a different order. Instead of generating a large number of candidates, depth-first search categories (e.g., PrefixSpan) take a more efficient approach which is focused on counting the frequency of the relevant data sets instead of the candidate sets (Dong & Pei, 2007). They scan the entire database to match against the whole set of candidates in each pass, and then partition the data set to be examined as well as the set of patterns to be examined by database projection (Slimani & Lazzez, 2013). Such a divide-and-conquer methodology substantially reduces the search space and leads to high performance (Dong & Pei, 2007).

As my discussion above shows different algorithms utilize different strategies to search for sequential patterns efficiently (Zhao & Bhowmick, 2003), they differ in the type of database representation they use, how generators determine the next patterns to be explored in the search space, and how they count the support of patterns to determine if they satisfy the minimum support constraints. Despite the differences, all sequential pattern mining algorithms return the same set of sequential patterns if they are run with the same parameter on the same database. Therefore, the difference between the various algorithms is not their output, but rather how each algorithm discovers the sequential patterns (Founier-Vinger et al., 2016).

Although sequential pattern mining is very useful in discovering common sequential patterns, it has its limitations. An important limitation of this technique is that it cannot assess the probability of a pattern followed by another pattern. To address this limitation, data mining scientist have developed other *sequential rule mining techniques* that account for the probability that a pattern will be followed (Founier-Vinger et al., 2016). A sequential rule is a rule of the form X -> Y where X and Y are sets of items. A rule X -> Y is interpreted as if items in X occur, then it will be followed by the items in Y. To find sequential rules, two measures are generally used: 1) support; and 2) confidence. The support of a rule X -> Y is how many sequences contains the items from X followed by the items from Y (Founier-Vinger et al., 2016). The confidence of the rule is the support of the rule divided by the number of sequences containing the items from X (Founier-Vinger et al., 2016). It can be understood as the conditional probabilities P(Y|X), expressed as Equations 1 and 2.

$$\text{Support (X -> Y)} = \text{support(X\&Y)} = Pr\ (X\ \&\ Y) \tag{1}$$

$$\text{Confidence (X -> Y)} = \frac{\Pr(X\ \& Y)}{\Pr(X)} = Pr\ (Y|X) \tag{2}$$

A sequential rule mining algorithm provides all sequential rules that have a support that are no less than threshold minimum (i.e., *minsup*) set by the researcher. To reduce the chance of losing any interesting course-taking pattern, especially broken sequences, I decided to set the minimum support value to the lowest place that algorithm would converge, which was 4 percent.

Many software packages have been developed to execute sequential data mining. For the purposes of this study, I utilized TraMineR for applying data mining tasks. TraMineR (**Tra**jectory **Min**er in **R**) is a R-package for mining, describing and visualizing discrete sequence data, especially designed for social science (Gabadinho, Ritschard, Müller, & Studer, 2009). I chose TraMineR since it is developed in R and, therefore, it has the advantage of its powerful graphical capacities. It is also a free source and its functions could be used in combination with R's other packages. The algorithm implemented in TraMineR is an adaptation of the Prefix-Tree-Based search, which is considered a depth-based search algorithm (Ritschard et al., 2012).

Procedures. I explored sequential patterns of majoring and course-taking in two separate analyses: 1) trajectories in students' academic majors; 2) sequential patterns in course-taking.

Trajectories in students' academic majors. In my analysis, I treated students' academic majors as a state sequence that established a profile of each student's declared academic major for each semester – that is, each profile represents the sequence of a student's major across 12 semesters. Student majors were coded as: 1) EN (engineering); 2) MA (mathematics); 3) PH (physical sciences); 4) CS (computer science); 5) LF (life science); 6) NS (non-STEM); and 7) UN (undeclared).

As a starting point, I looked at the distribution of students' majors in each semester and generated plots to visually-represent patterns. In particular, I was interested in understanding how pattern frequency in one academic major was related to pattern frequency in other academic majors, as well as to changes in students' major declaration (switching academic majors) and the point in time that students dropped out of the University. I used TraMineR to calculate transition rates between states, and to compare switching rates from STEM fields to non-STEM and also movement among majors within STEM.

Next, and since I was interested in exploring female and male differences in majors, particularly within STEM fields, I calculated a gender covariate to explore how female and male students' major distribution patterns differ. Most studies of gender differences in STEM fields (e.g., Chen, 2013; Griffith, 2010) looked at differences in enrollment or degree completion and have not been able to follow differences in major patterns from the beginning of students' study until the finishing point. They have also failed to look at the differences within STEM fields. Integrating a gender covariate into the analysis allowed me to investigate whether female and male students follow different paths when declaring academic majors.

Finally, I used cluster pattern analysis to build a typology of student major sequences. This allows me to identify groups of students with similar patterns in academic majors over time. To build such a typology, a clustering method is applied to aggregate the sequences into a reduced number of groups by measuring how alike two sequences are with each other. Clustering is an exploratory data analysis method aimed at finding automatically homogeneous groups or clusters in the data. For the purpose of this study, I will use the Hierarchical Ward clustering method (Ritschard, 2018), recommended by the software to cluster students' majoring patterns. Although it is difficult to provide a clear-cut solution about the "best" number of clusters in the data, a dendrogram plot provided by hierarchical clustering helps assessing the number of clusters by cutting a dendrogram at a certain level (Gabadinho et al., 2009). A six clusters solution was retained after examining the dendrogram plot of the clustering tree provided by Ward clustering method (see Figure 3.1). Once I identified the clusters, I ran a distribution analysis for each cluster to identify the most typical patterns that characterized the cluster. This analysis, with R's unique visualization, feature shows the distribution of academic majors that belong to each group. It also helps to identify which groups of students belong to each major cluster.

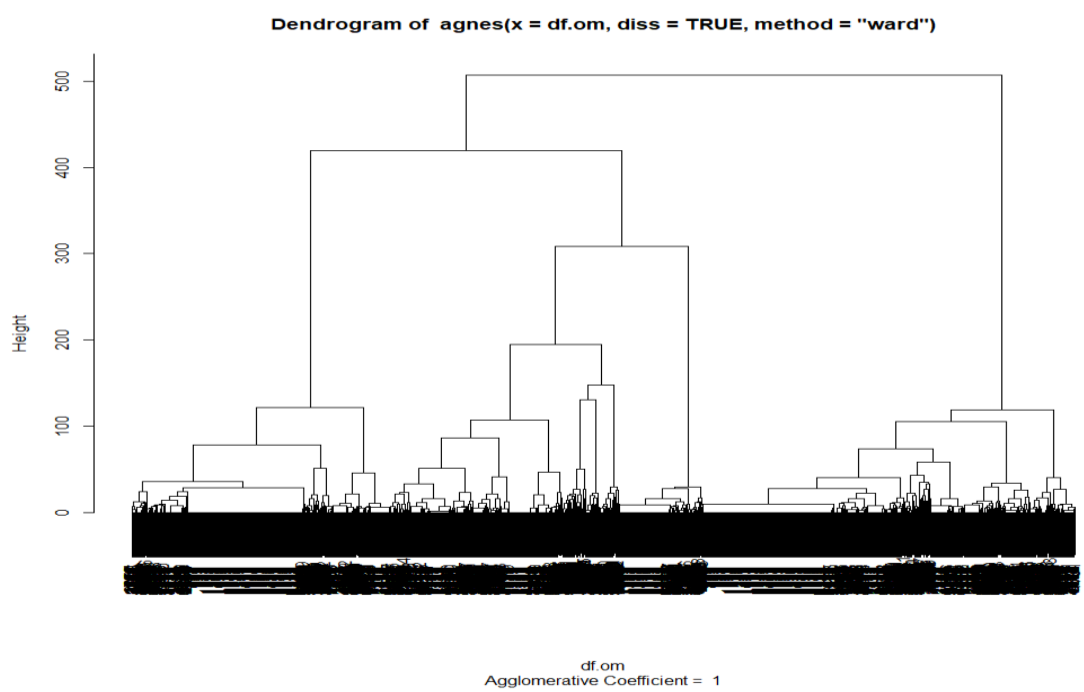

Figure 3.1. Hierarchical sequence clustering from the OM distances, Ward method

Course taking patterns. To examine course taking patterns I analyzed the sequences of events that identify frequent course taking patterns. To do so, I created an academic event profile for each student. Each itemset in the sequence represents STEM courses a student took in a particular specific semester – e.g., similar to a course transaction record for each semester. For example, during the first semester "student 1" took three STEM courses – Calculus I, Stat I, and Physics I – and two non-stem courses (see Figure 3.2). I do not include the STEM courses in my event since the focus of my analysis is on STEM course-taking patterns; in the second semester, he took Calculus II, Statistics II, Physic II and Computer Programming courses (see Figure 3.3).

As a first step in my analysis, I applied Sequential Pattern Mining to find frequent course-taking subsequences. To reduce the chance of losing any interesting course-taking pattern, especially sequences that leads to switching to non-STEM fields or quitting the University, I decided to set the minimum support value to the lowest place that algorithm would converge. By setting *minsupport* threshold to 4 percent only the patterns that appear in more than 4 percent of sequences are included in the search for frequent patterns. Sequential Pattern Mining finds the most common subsequence of course-taking pattern among students who took STEM courses. The results also provide information on the number of sequences that contain such a subsequence. I then plotted the results to visualize most frequent course-taking patterns.

Student_id		sex	semester	Courses
1	1	male	201201	Calculus I, Stat I, Physic I
2	1	male	201202	Calculus II, Stat II, Physic II, Programming
3	1	male	201301	Adv Math, Adv Engineering, Chemistry I
4	1	male	201302	Adv Engineering, Chemistry II, Physics I,
5	2	female	201202	Biology I; Stat I, Chemistry I
6	2	female	201203	Biology II, Chemistry II,

Figure 3.2 Sample Itemset Sequence for Hypothetical Students

The course sequence for the student-id #1 is:

{(Calculus I, Stat I, Physic I), (Calculus II, Stat II, Physic II, Programming),

(Adv Math, Adv Engineering, Chemistry I), (Engineering, Chemistry II, Physics I)}

The course sequence for students ID#2 is:

{(Biology I, Stat I, Chemistry I), (Biology II, Chemistry II)}

Figure 3.3 Sample Itemset Sequence for Hypothetical Students

Since I was most interested in courses that occur prior to switching between STEM and non-STEM majors or even dropping out of college, I had to identify broken sequences. To do so, I added a semester to the end of each student's profiles and looked for instances where patterns changed. For example, if a student's profile had eight semesters, I added a 9th semester to his/her profile. Similarly, for a profile with only three semesters, the last semester now was the fourth semester. This extra semester was coded as 'exit' for leaving the college. I added this to the end of all student sequences, which means that all students left after their last semester. Then, I dropped this semester from the profile of students who have completed their study in STEM fields. Next, I recoded this new course to 'NOSTEM' for students whose enrollment major was a non-STEM field. These were students whose majors required them to take STEM field courses or were considering switching to STEM fields but decided to stay in non-STEM field. Adding 'NOSTEM' to their sequence simply identifies the fact that their broken sequence does not mean they quit or switched to non-STEM fields. For students who started with a STEM field but completed their study in a non-STEM field, I recoded the new course to 'SWITCH'. This approach helped me to identify course-taking patterns for students who initially declared a STEM major, but who later switched to a non-STEM or left the university. In this way, by examining these broken sequences, I was able to identify potential "gate keeping" courses that were taken by students prior either switching majors or leaving college altogether.

To answer my second research question – i.e., whether female students take different course-taking patterns than male students – I identified the course-taking patterns that were most strongly related with female students. To do so, I used TraMineR's functions to execute discriminatory analysis. The results are subsequences, ordered by decreasing the discriminant power. I then measured the strength of association of each subsequence with the considered covariate and subsequently selecting the subsequences with the strongest association. The association was measured with the Pearson independence Chi-square. I use this function to find which sequence patterns best categorizes women.

Although sequential pattern mining does provide important information on the most frequent course-taking subsequences and the number of students whose academic profiles contain those subsequences, there is no assessment of the probability that an event will be followed by another event. To address this problem, I used sequential rule mining to discover sequential rules in students' course-taking sequences. Such rules provide insights into sequential patterns since they give a measure of confidence for their occurrence. Since my sequential pattern mining results show that there are specific courses present in quitters' and switchers' profiles, my goal was to estimate the probability of quitting or switching to non-STEM fields for students with particular course taking sequences.

CHAPTER 4: FINDINGS

The study's findings are presented in three parts. First, I describe the share of students with declared majors when they first matriculated as a degree seeking student at the University. I then explore how this initial distribution of students' academic majors changes over time, and the extent to which patterns differ for female and male students. To do so, I employ two different data mining techniques – Sequential Pattern Mining and cluster analysis of academic major sequences. Both techniques provide a somewhat different perspective on students' trajectories, as well as a useful comparison among potential methods for exploring patterns in higher education students' academic major trajectories.

In the third section, I explore differences in course taking patterns – for all students, and separately, for female and male students. This study uses a longitudinal approach to identify course-taking patterns. Most previous studies (e.g., Bahr, 2013; Wang, 2016) considered the number of STEM courses taken over time, overlooking the variations in the sequence of course-taking by the students as they progress along their college pathways. Using Sequential Pattern Mining techniques, I am able to identify the most frequent course-taking patterns in STEM – considering students' profiles. It also helps to discover the most discriminant course taking patterns between male and female STEM-considering students. Finally, Sequential Pattern Mining is employed to discover courses/sequence of courses that taking them lead to switching to a non-STEM major or dropping out of the University.

Distribution of Student Majors at Point of Matriculation

Total sample. Table 4.1 describes the distribution of academic program majors *for all students* who matriculated to the University during the Fall semester 2010, 2011 and 2012. Student academic majors represent those declared by students during their first semester as a matriculated student.

Altogether, about 30% of students declared a major in a STEM field, whereas about half of students declared a non-STEM major (51%). Among STEM majors, 19% of students declared a major in life sciences and about 7% declared a major in engineering. About 1% of students declared a major in computer science and mathematics (respectively), and 1.7% declared a major in physical science. For students in the three cohorts included in this study, 19% did not declare a major at their point of entry to the University.

There were notable differences in majors between male and female students. Male students were more likely to declare a major in a STEM field – i.e., 35% vs. 26% (male vs. female). In contrast, the majority of women (56%) declared a major in a non-STEM field, compared to 45% of male students. There were also differences among students who declared a STEM major. Women were more likely to declare a major in life science than their male peers (22 % vs. 15.2%) and less likely to declare an engineering major (2.5% vs. 13.5%). Although for both groups the share of students who declared a major in computer science was small, women were less likely to do so than men (0.2% vs. 2%). The same was true for physical science – i.e., 1.0% vs. 2.5% (women vs. men).

Table 4. 1

Student Academic Major Declared During First Semester as Matriculated Students

Academic Major at Point of Entry	Male	Female	Total
STEM Majors:	34.6	26.3	29.9
Computer science	2.2	0.2	1.1
Engineering	13.5	2.5	7.3
Life Sciences	15.2	22.0	19.0
Mathematics	1.2	0.5	0.8
Physical Science	2.5	1.1	1.7
Non-STEM Majors	45.4	56.1	51.4
Undeclared	20.0	17.7	18.7
Total	4,003	5,083	9,086

Note: Statistics are reported for the population of students who matriculated to the University of Vermont during the Fall 2010, 2011, and 2012 semesters.

STEM-considering students. Table 4.2 describes the distribution of academic majors for the subset of students who took more than two STEM-related courses during their first year of study at the University. As discussed in Chapter 3, I refer to this group of students as “STEM-considering” based on their initial course taking pattern. STEM-considering students include those with and without declared majors in STEM, since students with non-STEM majors or undeclared majors may have enrolled in STEM-related courses during their first year of study. Altogether, 4,890 students matriculating to the University during 2010, 2011 and 2012 semesters were STEM-considering. This is equivalent to about 53% of Fall and Spring semesters matriculating students.

The majority of STEM-considering students declared a STEM major (55%). However, about 30% of students who took more than two STEM courses during their first year were non-STEM majors, and 14.8% were undeclared majors. Interestingly, female STEM-considering students were more likely to declare a non-STEM major than their male counterparts (36.5% vs. 21.9, female vs. male). Conversely, male STEM-considering students were more likely to not declare a major at their point of entry to the University than females (12.9% vs. 17.2%, female vs. male).

Table 4. 2

Academic Major Declared in First Semester for STEM-Considering Students

Major in Enrolment	Male	Female	Total
STEM Majors:	60.93	50.61	55.40
Computer science	3.84	0.46	2.03
Engineering	23.91	4.73	13.61
Life Science	26.82	42.33	35.15
Mathematics	2.12	1.03	1.54
Physics	4.24	2.06	3.07
Non-STEM Majors	21.92	36.54	29.76
Undeclared	17.15	12.85	14.84
Total	2,265	2,625	4,890

Graduation rates across academic majors. Table 4.3 presents academic majors, at point of entry and also degree conferred at graduation for STEM-considering students – that is, the table presents the percentage of STEM-considering students across academic majors in their first semester and when students graduated from the University.

Overall, more than half of STEM-considering students initially declared a STEM-related major (55.4%). However, the share of students who actually complete a degree in STEM related field is substantially less, just 38.6% of students. This is equivalent to about a 30% decrease between students' first and last semesters. The rate of decline in the share of STEM majors between point of entry and graduation is about the same for male and females – however, there were fewer female students in STEM majors to start with.

Among STEM-considering students, students who declared a STEM major at the point of entry were most likely to declare a life science major (35.2%) compared to other STEM majors; however, just 23.7% graduate with this major. That is to say, about one-third of students who enter the University declaring a life science major did not graduate with this major. By comparison, female students who initially declared a life science major were more likely to persist with this major through graduation – i.e., 42.3% of STEM-considering female students initially declared a life science major, and 29.9% graduated with this major. Whereas for men, just 26.8% of STEM-considering students declared a life science major, and only 16.5% graduated with a degree in this major (a decline of 39.5%).

Similarly, while 13.6% of STEM-considering students initially declare an engineering major, just about 71% complete their degree in engineering (9.7% of all

STEM-considering students). In contrast, over time, the share of STEM-considering students with a non-STEM degree grows as students continue their progress toward graduation. The share of men and women who initially declare an engineering major, and who then complete an engineering degree, is about 70% for both groups. That said, the share of women who pursue an engineering degree is considerably less than their male peers. Initially, 29.8% of the sample declared a major in a non-STEM field, and subsequently 39% of students graduated with a non-STEM major.

Table 4. 3

Academic Major Declared in First Semester and Degree Conferred at Graduation for STEM-Considering Students

	Male		Female		Total	
Academic Major	Enrollment	Graduation	Enrollment	Graduation	Enrollment	Graduation
STEM Majors	60.9	41.5	50.6	35.3	55.4	38.6
Computer science	3.8	3.2	0.5	0.5	2.0	1.7
Engineering	23.9	17.0	4.7	3.3	13.6	9.7
Life Science	26.8	16.5	42.3	29.9	35.2	23.7
Mathematics	2.1	2.9	1.0	1.6	1.5	2.2
Physical Science	4.2	1.9	2.1	0.7	3.1	1.3
Non-STEM	21.9	33.4	36.5	43.9	29.8	39.0
Undeclared	17.2		12.9		14.8	
Incomplete		25.2		20.1		22.4
Total Students	2,265		2,625		4,890	

Student Academic Majors Overtime

Figure 4.1 provides a visual summary of STEM-considering students' academic major trajectories over time. It provides a broad overview of majoring patterns, their frequencies, how they compare to each other, and how such frequencies relate to switching and/or dropping out of the University. This figure covers all 4,890 students in my analysis. The sequences represent their term majors for each semester over the course of six years. The x-axis is the semester and the y-axis show the sequences' accumulated frequency in percentage of students. Students' majors have been recoded to seven categories: non-STEM, life sciences, engineering, physical science, mathematics, computer science, and undeclared respectively denoted as NS, LF, EN, PH, MA, CS, UN. Each of these codes are represented with a specific color on the plot. The green represents missing majors, meaning that the student did not have any major record, that is to say, the student is not enrolled in the University anymore.

As it is clear from the plot, life science majors dominate majoring patterns within STEM fields at the point of entry, attracting more students compared to other STEM majors. However, the frequency of this major declines as students go further in their study. That is to say, students who enter the University declaring a major in life science later switch to other majors, specifically to non-STEM fields. The second most frequent pattern in STEM fields is the engineering path. While in comparison to life science, engineering has a much lower frequency, even this low frequency declines as the students go further in their course of study. Non-STEM majors' domination, in contrast, grows as students continue their study towards graduation, meaning that students are switching from STEM or undeclared majors to non-STEM majors. What this figure suggests is that there is a dynamic process at work for how student trajectories develop over 12 semesters.

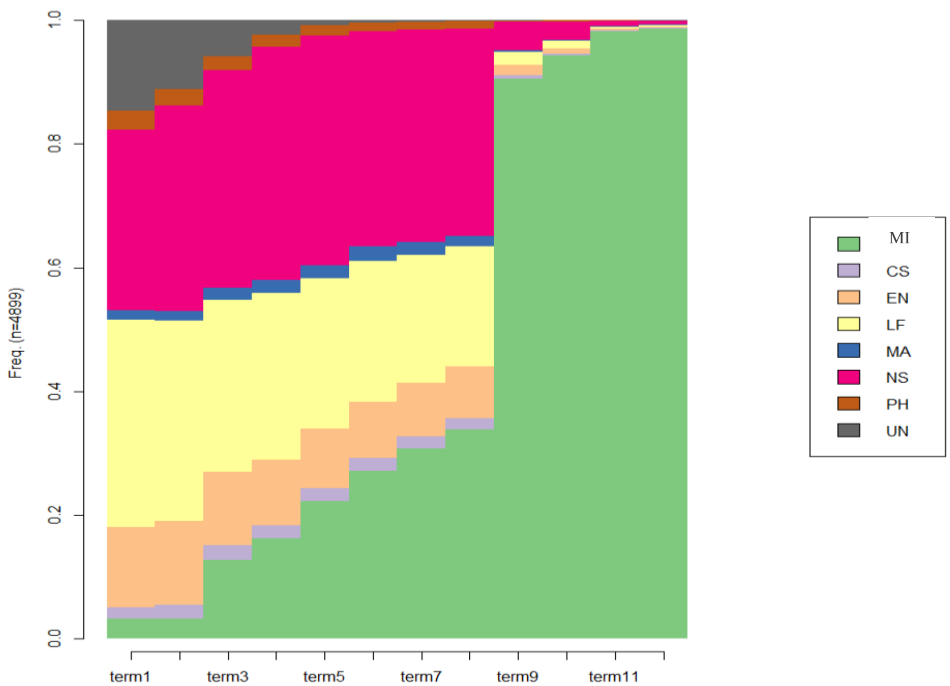

Figure 4. 1 Visual Summary of Patterns in Academic Majors for STEM-Considering Students (Over 12 Semesters)

Note: The plot represents STEM-considering student majors for each semester over the course of six years (12 semesters). The x-axis is the semester and the y-axis show the sequences' accumulated frequency in percentage of students.

In the following sections, I will explore this dynamic process with more depth using Sequential Pattern Mining and Cluster Pattern Analysis. Using Sequential Pattern Mining, I identify the most frequent patterns for academic majors over students' enrollment period. This helps us understand how and when students change academic majors. By comparison, cluster analysis helps to create a student typology according to their academic major sequences and analyze group differences within these clusters, particularly differences between men and women.

Pattern analysis. In this first section, I present findings from my sequential pattern analysis with STEM-considering students. Specifically, I consider: 1) dominant academic major patterns, for the overall sample and by gender; and 2) students' transitions among academic majors.

Dominant academic major patterns. Table 4.4 describes the 10 dominant academic major patterns for all STEM-considering students identified by Sequential Pattern Mining. The 10 dominant patterns cover 43.9% of STEM-considering students in the three cohorts included in this study – put another way, this means that more than half of students pursued other pathways that did not necessarily conform to some overall trend in academic major selection.

The first key finding is that nearly one-third of STEM-considering students start and complete the same academic major within four years. Specifically, about 14% students who initially declared a non-STEM major persisted in a non-STEM major for eight semesters. This percentage represents slightly less than half of the students who initially declared a non-STEM major (29.8%, Table 4.3). Among STEM fields, 12.3% of students who initially declared a life science major completed their major in eight semesters; this was just about one-third of the students who initially declared a life science major (35.2%, Table 4.3). By comparison, 5.3% of students who initially declared an engineering major persisted with this major for eight semesters, or about 60%

of students who initially declared an engineering major (13.6%, Table 4.3). Taken together, these findings suggest that among students initially declaring a STEM major, sizable shares of students are not completing that major in four years.

Interestingly, the second set of frequent patterns that emerged were for students who dropped out of the University after two semesters. About 2% of students who started in a non-STEM major dropped out spring of their freshman year, and another 2% of life science majors dropped out then as well. Finally, about 1.5% of students who were initially undeclared majors switched to a non-STEM major after two semesters, and subsequently persisted with a non-STEM major for another six semesters.

Table 4. 4

Ten Most Frequent Patterns in Academic Majors among STEM-Considering Students

Subsequence	**Frequency**	**Percentage**
Non-STEM / 8 Semester	650	13.7
Life Science / 8 Semester	583	12.3
Engineering / 8 Semester	250	5.3
Non-STEM / 2 Semester	105	2.2
Life Science / 2 Semester	99	2.1
Life Science / 6 Semester	97	2.0
Non-STEM / 6 Semester	86	1.8
Non-STEM/ 5 Semester- Non-STEM/2 Semester	77	1.6
Undeclared / 2 Semester- Non-STEM /6 Semester	70	1.5
Life Science / 4 semester	68	1.4
Total students	2,085	43.9

Note: This table lists the 10 most frequent patterns identified using Sequential Pattern Mining for academic majors among *all* STEM-considering students.

Differences in dominant academic major patterns for female and male students. I also find differences in the academic major patterns for female and male STEM-considering students. The visual comparison of female vs. male majoring pattern distribution (Figure 4.2) clearly shows significant differences between male and female students' majoring patterns overtime. As the comparison plot shows, the majority of female students who enroll in STEM fields follow a trajectory in life science and very few of them follow an engineering path. In contrast, male students follow life science and engineering paths at the same rate. Life science and engineering paths also decline for both male and female student as they go further in their studies. The male-female differential, however, remains significant.

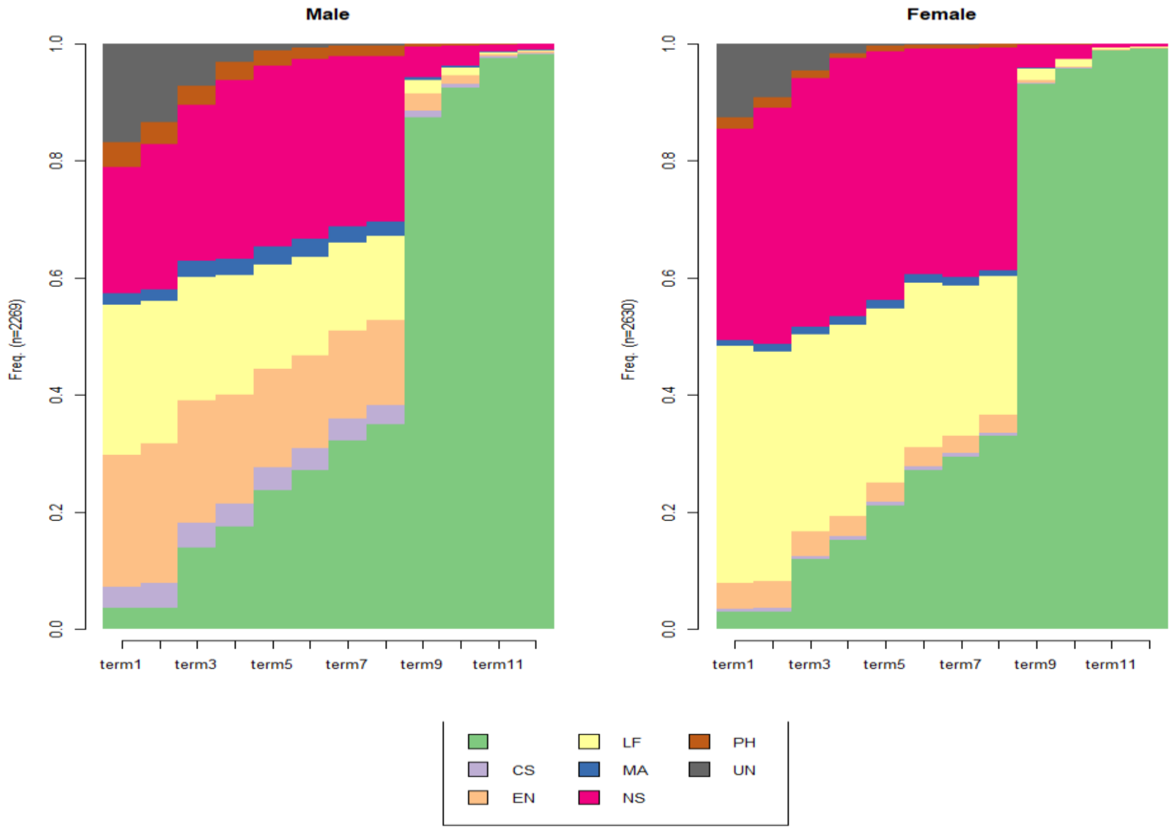

Figure 4. 2 Gender Differences in Academic Majors over Time

Note: The plot represents STEM-considering student majors for each semester over the course of six years (12 semesters). The x-axis is the semester and the y-axis shows the sequences' accumulated frequency in percentage of students. Majors are denoted as NS: non-STEM, LF: life science, EN: engineering, PH: physical science, MA: mathematics, CS: computer science, UN: undeclared

As was the case above, I focused on the 10 most dominant academic major patterns for STEM-considering students (Tables 4.5 and 4.6). Overall, about 28% of male STEM-considering students completed within four years the academic major they initially declared upon entering the University (see Table 4.4). As was the case for the full sample, the three most dominant patterns for male students were for students who initially declared a non-STEM major, and among STEM majors those that initially declared an engineering or life science major. For male non-STEM majors, less than half persisted with their major for eight semesters (9.6% vs, 21.9%). A similar attrition rate was apparent for male students with initially-declared STEM majors – for engineering, 9.3% of students persisted for eight semesters (of 23.9% who initially declared); and for life science, 9.1% persisted (of 26.8% who initially declared).

For females, 32% persisted with their initial academic major for eight semesters – however, these majors were limited to non-STEM and life science. There was no similar pattern among females for engineering; that is, female persistence in an engineering degree was the eighth dominant pattern for academic majors. For females, 17.1% who initially declared a non-STEM major persisted for eight semesters; this is slightly more than half of the women who initially declared a non-STEM major. However, for women who initially declared a life science major, just about one-third of those who initially declared persisted in this major for eight semesters (i.e., 14.9% of 42.3% who initially declared), and another 2.7% of females who initially declared as a life science major persisted for six semesters.

The pattern for student dropouts differed for males and females. For males, it was a dominant pattern, with nearly 5% of the sample dropping out of two semesters – 2% of which were initially non-STEM majors, 1.7% were initially life science majors, and 1.6% were engineering majors. About 5% of females also dropped out after two semesters, but their initial academic majors were somewhat different, with 2.4% initially declaring a life science major and the other 2.4% a non-STEM major.

Among male STEM-considering students, about 5% of the sample switched from an undeclared major to a non-STEM major after their second or third semester. There was no similar pattern for female students. However, 1.6% of females switched from life science to non-STEM after two semesters.

Table 4. 5

Ten Most Frequent Patterns in Academic Majors among STEM-Considering Students: Males

Patterns	**Frequency**	**Percentage**
Non-STEM / 8 Semester	211	9.6
Engineering / 8 Semester	204	9.3
Life Science/ 8 Semester	201	9.1
Non-STEM / 2 Semester	43	2.0
Life Science / 2 Semester	37	1.7
Engineering /2 Semester	36	1.6
Non-STEM/6 Semester	36	1.6
Undeclared /2 Semester-Non-STEM/6 Semester	36	1.6
Undeclared /3 Semester-Non-STEM /5 Semester	30	1.4
Life Science /6 Semester	29	1.3
Total Students	863	37.6

Note: This table lists the 10 most frequent patterns identified using Sequential Pattern Mining for academic majors among *males* STEM-considering students.

Table 4. 6

Ten Most Frequent Patterns in Academic Majors among STEM-Considering Students: Females

Patterns	**Frequency**	**Percentage**
Non-STEM /8 Semester	439	17.1
Life Science /8 Semester	382	14.9
Life Science /6 Semester	68	2.7
Non-STEM /5 Semester- Unenrolled /1 Semester-Non-STEM/ 2 Semester	66	2.6
Life Science / 2 Semester	62	2.4
Non-STEM /2 Semester	62	2.4
Non-Science /6 Semester	50	2.0
Engineering /8 Semester	46	1.8
Life Science /7 Semester	43	1.7
Life Science /2 Semester-Non-STEM /6 Semester	41	1.6
Total Students	1.259	49.2

Note: This table lists the 10 most frequent patterns identified using Sequential Pattern Mining for academic majors among *female* STEM-considering students.

Switching patterns among academic majors. My distribution analysis of most frequent academic patterns demonstrated that not only a lower number of students begin their college career with a STEM major, but also those numbers decline over time as a result of some leaving STEM for other fields. To arrive at a better understanding of the switching patterns among academic majors, with particular attention to students switching between STEM and non-STEM majors and within STEM fields, I built three transition matrices – one for all STEM-considering students; and two others, for female and male STEM-considering students separately. When considering switching between academic majors, I look at any changes during the 12 semesters after initial enrollment.

Table 4.7 summarizes the primary switching patterns that emerged from the data. Since students who majored in engineering, mathematics, physical, or computer science comprised a small share of STEM-considering students (10.9%), for this analysis I combined students declaring one of these majors into a new general category titled “hard sciences.”

Overall, I find that the share of students switching from STEM to non-STEM majors is higher than the share of students switching from non-STEM majors to STEM majors (18.7% vs. 1.5%). Specifically, 14.8% of students who initially declared a major in the “hard sciences” switched to a non-STEM major sometime over the course of the next 12 semesters, and another 5.2% switched from a major in the hard sciences to a life science major. Conversely, a very small share of students switched from a non-STEM or a life science major to a major in the hard sciences (about 0.5%, respectively). Similarly, the share of students who switched from non-STEM to STEM majors was small, just 1% of non-STEM students switched to a life science major and 0.5% to a major in the hard sciences.

Switching patterns, however, were considerably different for female and male students. Specifically, females were more likely to switch from a major in the hard sciences to a non-STEM field, and within STEM majors from the hard sciences to life sciences. Female students who majored in the hard sciences left for non-STEM majors at a much higher rate than their male peers – 17.7% vs. 13.0% (female vs. male). This is notable given the comparatively small number of females who declared a major in the hard sciences. When females switched from the hard sciences to another STEM field, they were more likely to declare a life science major than their male counterparts (10.3% vs. 3.2%). Female and male students who initially declared a life science major were equally likely to switch to a non-STEM major (4% vs. 3.8%). These findings are particularly notable given the comparatively small number of females who initially declared a major in the hard sciences; that is, females switch to non-STEM fields at higher rates and the very few who stay in STEM were more likely to move from academic majors in hard science to a life science major.

Table 4. 7

Transitions among Academic Majors over 12 Semesters

Transition	All students	Female	Male
Non-STEM => Life Science	1%	1.1%	0.6%
Non-STEM => Hard Science	0.5%	0.3%	0.6%
Life Science => Hard Science	0.5%	0.4%	1.2%
Life Science => Non-STEM	3.9%	4%	3.8%
Hard Science => Life Science	5.2%	10.3%	3.2%
Hard Science => Non-STEM	14.8%	17.7%	13.1%

Note: Hard Science include engineering, computer science, Physical Science, and Mathematics majors.

The findings from Sequential Pattern Mining of students' academic majors suggest that among STEM-considering students who declared a STEM major upon entry, a sizable share did complete their major in four years. Also, the findings show that a large number of students pursued a pathway that did not necessarily conform to any of the overall trends in academic major selection discussed above. Within STEM, life science majoring patterns were the most frequent among STEM-considering students. The frequency of such patterns, however, declines over time as students switch to non-STEM majors. Even though engineering and other hard science paths are much less frequent, they follow a similar pattern of decline as students go through further in their study. The findings also reveal that male and female students follow clearly different academic paths and that this gender-based difference becomes even more significant within STEM fields. That is to say, more female STEM-considering students follow non-STEM paths and the number of such students grows as they continue their studies. Within STEM fields, life science trajectories enjoy a much higher level of popularity among female students. In contrast, engineering is much more popular (the second most frequent pattern) among male STEM-considering students. Despite these important differences, the popularity of STEM paths declines for both female and male students as they further progress in their studies. Transition analysis confirms that, in general, a higher number of students switch form STEM to non-STEM compared to the number of students switching otherwise. The rate of switching from STEM to non-STEM is even higher for female students. Within STEM fields, more students switch from a hard science major to non-STEM compared to students who switch from life science. Women comprise most of the switchers from hard science to non-STEM majors. All these patterns point out to the fact that the institution is struggling to recruit and keep students, especially women, in STEM fields, particularly in hard science.

Cluster Analysis. Another way to consider students' academic trajectories is to cluster students according to their academic major sequences. Specifically, I used this approach to better understand gender differences in academic majors. While Sequential Pattern Mining provides the most frequent sequences in academic majors, cluster analysis allows us to look within similar groups of students (according to academic major) to better understand different decision-making patterns. This allows me to develop a typology of students based on academic major– similar to what Adelman (2005) and Bahr (2010) did in earlier research. However, I build on these earlier works to take into account sequencing in academic major when clustering students, rather than just clustering students based on their academic majors at one point in time.

Cluster analysis identified six student groups with academic majors in: 1) life science (Cluster 1); 2) physical science and mathematics (Cluster 4); 3) engineering (Cluster 5); 5) computer science (Cluster 6); and 6) non-STEM fields (Cluster 3). The procedure also identified a distinct group of students who dropped out of the University sometime between their first and twelfth semester enrolled (Cluster 2). Figure 4.4 depicts the distribution of academic majors in each cluster of students that occurred over 12 semesters.

Table 4.8 shows the percentage of female and male students who belong to each group and Table 4.9 shows the most frequent majoring patterns for each cluster. In what follows, I will point out to some of the significant findings that we can derive from each cluster, and the two associated tables. Table 4.9 shows that around half of students in life science cluster start their study in a major in life science and persist in it for seven or eight semesters. Table 4.8 shows that life sciences are dominated by women, 16% of all female students compared to 9% of males. The second cluster, titled “the Quitters” represents sequences in which students drops out of the University after a few semesters. Based on Table 4.8, 17.5% of the quitters are life science majors who left the University after their first or second semester, while 10.2% of them are engineering majors who dropped out after the first or the second semester (and sometime even after their fourth semester). Around 4% of the Quitters are students who did not declare a major initially and left the University after two semesters without ever having selected a major. Table 4.8 shows that there is a slight gender disparity in this cluster. A total of 10% of male students are among the quitters while only 8% of female students belong to this category. The third cluster of Figure 4.3 represents non-STEM majors as well as the students who switched to non-STEM fields. Around 10% of students in this group did not declare a major when enrolled and then switched to a non-STEM major after their first, second, or sometimes even third semester of the study. It is important to note that this cluster is dominated by female students as well (24% of females vs 16% males). The last three clusters presented on Figure 4.3 are hard science major groups (engineering, physical science, and mathematics). An initial characteristic that all these three clusters share is the much lower number of students, compared to other clusters, that belong to them. Additionally, Table 4.8 shows that all these three clusters are dominated by male students, female students being significantly underrepresented in all, almost absent in computer science.

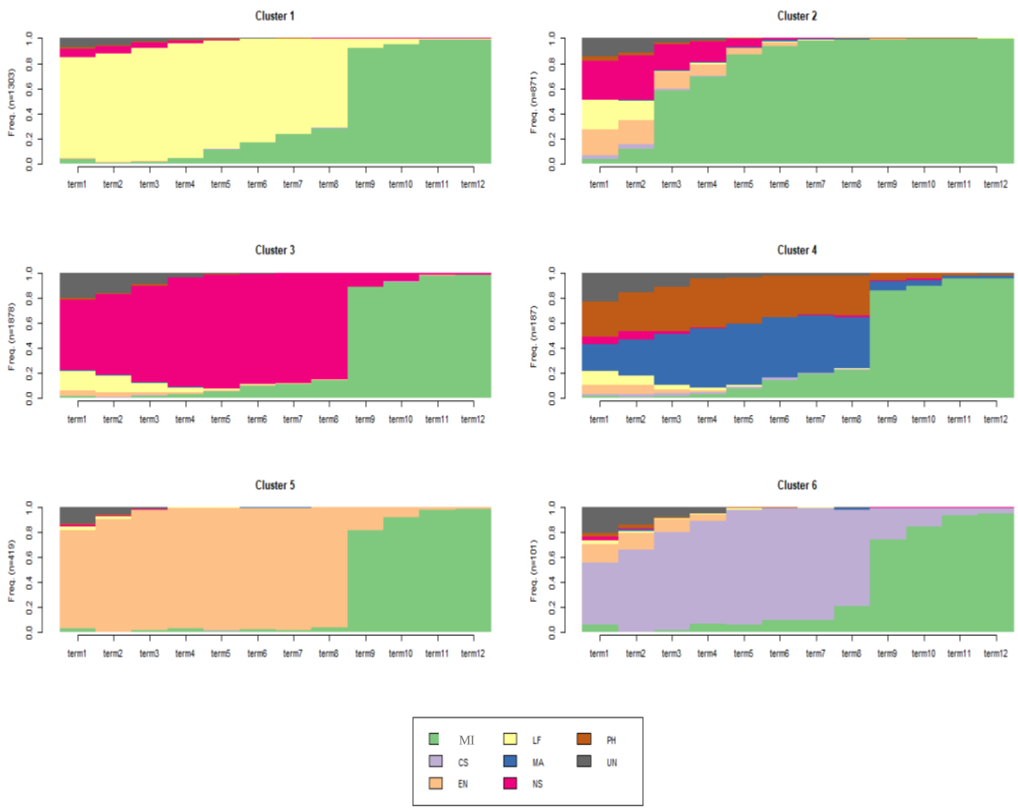

Figure 4. 3 Clusters of Students' Academic Major

Note: Majors are denoted as NS: non-STEM, LF: life science, EN: engineering, PH: physical science, MA: mathematics, CS: computer science, UN: undeclared.

Table 4. 8

Academic Major Clusters with Male and Female Students Membership

Cluster	**Male**	**Female**
Cluster 1: Life Science	9%	18%
Cluster 2: Quitters	10%	8%
Cluster 3: Non-STEM	16%	24%
Cluster 4: Physics and Mathematics	3%	1%
Cluster 5: Engineering	7%	2%
Cluster 6: Computer Science	2%	0%

Table 4. 9

Frequent Major Patterns for Each Academic Major Cluster

sequences	Counts	percent	sequences	counts	percent
Life Science			Quitters		
LF/8 Semester	583	44.7	NS/2 Semester	105	12.1
LF/6 Semester	97	7.4	LF/2 Semester	99	11.4
LF/4 Semester	68	5.2	LF/1 Semester	53	6.1
LF/7 Semester	55	4.2	NS/4 Semester	49	5.6
LF/5 Semester-/1-LF/2 Semester	32	2.5	EN/2 Semester	43	4.9
LF/5 Semester	27	2.1	UN/2 Semester	34	3.9
LF/9 Semester	25	1.9	NS/3 Semester	33	3.8
LF/3 Semester	24	1.8	EN/4 Semester	27	3.1
Non-STEM			Physics/Math		
NS/8 Semester	650	34.6	PH/8 Semester	24	12.8
NS/6 Semester	86	4.6	MA/8 Semester	15	8
NS/5 Semester -/1-NS/2 Semester	77	4.1	EN/3 Semester -MA/5 Semester	4	2.1
UN/2 Semester -NS / 6 Semester	70	3.7	PH/6 Semester	4	2.1
UN/3 Semester -NS/5 Semester	55	2.9	UN/3 Semester -MA/5 Semester	4	2.1
LF/2 Semester -NS/6 Semester	54	2.9	LF/2 Semester -MA/6 Semester	3	1.6
UN/1 Semester -NS/7 Semester	50	2.7	LF/2 Semester -PH/6 Semester	3	1.6
NS/7 Semester	31	1.7	MA/4 Semester -/1-MA/3 Semester	3	1.6
Engineering			Computer Science		
EN/8 Semester	250	59.67	CS/8 Semester	21	20.8
UN/1 Semester -EN/7 Semester	21	5.01	CS/6 Semester	7	6.9
EN/9 Semester -/3 Semester	17	4.06	CS/9 Semester	5	5
UN/2 Semester -EN/6 Semester	15	3.58	UN/2 Semester -CS/6 Semester	4	4
EN/7 Semester	12	2.86	CS/7 Semester	3	3
/1-EN/7 Semester	10	2.39	EN/1 Semester -CS/7 Semester	3	3
EN/10 Semester	7	1.67	UN/1 Semester -CS/7 Semester	3	3
EN/3 Semester -/1-EN/4 Semester	5	1.19	/1-CS/ 5 Semester	2	2

Note: Majors are denoted as NS: non-STEM, LF: life science, EN: engineering, PH: physical science, MA: mathematics, CS: computer science, UN: undeclared.

To conclude, cluster analysis offers a way to group students together based on their academic behavior patterns over time without considering any other information related to student characteristics. In an ideal situation, there should be no strong association between non-academic student characteristics and membership in a cluster. That is to say, in an ideal institution we expect students from different racial, ethnic, or class backgrounds to be represented roughly equally in all clusters. The results from cluster analysis, however, clearly demonstrates that this is not the case. As I have shown, female students are over-represented in life science and non-STEM clusters and significantly under-represented in engineering and computer science fields. Such associations are clear indications of systemic problems that lead to an unlevel playing field in which groups of students, like male students, are positioned better compared to their female counterparts to pursue certain majors. Unfortunately, since I did not have access to other demographic information, such as information on student socioeconomic status or their pre-college records, I was unable to determine whether there are other characteristics that are strongly associated with certain clusters beyond gender.

Student Course Taking Over Time

A second purpose for this study was to understand how student course-taking experiences related to whether or not they completed a degree in a STEM-related field. Most previous studies that have considered course-taking patterns simply look at the number of STEM courses taken and the relationship between this number and degrees obtained (e.g., Chen, 2013, 2015). This study takes a different approach and looks at the actual sequence of courses taken by students and the likelihood that a student completes a degree in a STEM-related field. I accomplish this using Sequential Pattern Mining techniques. As a second step, I also developed a typology of students based on their course taking behavior using cluster analysis techniques. The resulting typology helps us to understand whether specific course taking behaviors are associated with gender – i.e., are certain course taking behaviors more likely for women or men.

Course taking patterns over time.

***Pattern analysis*.** In this section, I present findings from my Sequential Pattern Analysis for STEM-considering students. Specifically, I consider dominant patterns in sequential course taking by students, overall and by gender.

*Dominant patterns in sequential course taking***.** First, I examined the most frequent course taking patterns among STEM-considering students. The patterns represent the sequence in which courses were taken; however, it may be the case that the sequence occurs over multiple semesters and the patterns do not necessarily represent courses taken in sequential semesters. The purpose of this analysis was to identify "broken" sequences that identify course taking patterns that lead to students leaving the STEM fields. By broken sequences, I mean the point at which students stop taking STEM-related courses or leave the University altogether. I consider three patterns – 1) students who continue to take STEM-related courses and finish degree at the University; 2) "switchers" – i.e., by switchers I mean students who initially declared a STEM major, who then after a particular sequence of courses stop taking any STEM classes; and 3) "quitters," who initially declared a STEM major and then leave the University after taking a certain sequence of STEM-related courses.

Table 4.10 shows the most 12 frequent course-taking sequences, sorted by frequency (support). That is to say, 33% of all students chose to take chemistry and calculus concurrently. None of the 12 most frequent course-taking patterns identify sequences where students switch to non-STEM course taking, or “quitters”. The most frequent course taking sequence is Calculus I and then Calculus II, with about 40% of STEM-considering students completing that sequence. The second most frequent pattern was for students to take Calculus I and Chemistry I concurrently; about one-third of students followed this pattern. About 30% of students took Chemistry I and then Calculus II (29%), and another 28% took Chemistry I and then Chemistry II. About 26% of students took an Introductory Science course followed by another Introductory Science course. Finally, one-quarter students took Calculus I and Chemistry I concurrently and then followed up with Calculus II (25%).

Table 4. 10

Most Frequent Course-taking Pattern

Sequence	Support	Count
Calculus I → Calculus II	39.9%	1,900
(Calculus I, Chemistry I)*	33.4%	1,588
Chemistry I → Calculus II	28.7%	1,365
Chemistry I → Chemistry II	28.4%	1,352
Science Introductory → Science Introductory	26.0%	1,236
(Calculus I, Chemistry I)* → Calculus II	25.3%	1,202
Calculus I→ Science Introductory	23.4%	1,113
Calculus I → Science Advanced	23.3%	1,110
Calculus I → Statistics 141	23.2%	1,103
Chemistry I → Science Advanced	22.4%	1,067
Science Advanced → Science Advanced	21.6%	1,028
Calculus I → Chemistry II	20.6%	981

Note: Science introductory courses includes a range of introductory level science courses (<=100 level) that students might take early on in their academic careers. Science advanced courses are comprised of general science courses at the 200 level or above. * notes courses that are taken concurrently. The patterns represent the sequence in which courses were taken; however, it may be the case that the sequence occurs over multiple semesters and the patterns do not necessarily represent courses taken in sequential semesters.

Since I was most interested in identifying the course taking patterns that preceded an initially-declared STEM major to switch to a non-STEM major, I looked for sequences in course taking that occurred prior to switching to a non-STEM major. Table 4.11 lists the most common course taking patterns for students who switched majors or quit from the University after taking these courses. There are several notable patterns. First, it appears that course taking sequences that involve Calculus I, Calculus II, and Chemistry I occur more frequently among students who were initial STEM majors who then switch to non-STEM majors. Altogether, about 13% of students who initially declared a STEM major and who took Calculus course work switched to a non-STEM major. Specifically, about 5% of students who initially-declared a STEM major switched to a non-STEM major after taking Calculus I, and 8% of students who took Calculus I and then Calculus II later switched to a non-STEM major. This equates to about 603 students (over three cohorts) who were initial STEM majors that did not graduate with a STEM-related degree. Additionally, about 5% of students who took Chemistry I later switched to a non-STEM major (this equates to 206 students, across three cohorts), and 6% of students who took Calculus I and Chemistry I concurrently also switched to a non-STEM major (this equates to 285 students, across three cohorts). Interestingly, there were similar patterns among initial STEM majors who subsequently left the University. Six percent of students who took Chemistry I subsequently left the University, and 7.8% of students who took Chemistry I and Calculus I concurrently also left. About 6% of students who initially declared a STEM major left the University after taking Calculus I, and 6.3% of students who took Calculus I and then Calculus II left.

Taken together, these findings suggest that the introductory Calculus sequence and Chemistry I are pivotal courses for whether students continue to pursue a STEM-related degree or leave the University altogether.

Table 4. 11

Course-Taking Patterns for STEM Majors Who Subsequently Switched Majors or Dropped Out of University

Sequence	Support	Count
(Calculus I) ⟶ (Calculus II) ⟶ (Switch)	7.7%	367
(Calculus I) ⟶ (Switch)	5.0%	236
(Calculus I) ⟶ (Calculus II) ⟶ (Quit)	6.3%	299
(Calculus I) ⟶ (Quit)	5.3%	270
(Chemistry I) ⟶ (Switch)	4.5%	206
(Calculus I, Chemistry I)* ⟶ (Switch)	6.0%	285
(Chemistry I) ⟶ (Quit)	5.9%	280
(Calculus I, Chemistry I)* ⟶ (Quit)	7.8%	372

Note: "SWITCH" identifies course taking patterns for students who initially declared a STEM major, and then after a particular sequence of courses stop taking any STEM classes. "QUIT" identifies course taking patterns for students who initially declared a STEM major and then left the University after taking a certain sequence of STEM-related courses. A complete list of full course titles alongside their designated code appears in Appendix C. * notes courses that are taken concurrently. The patterns represent the sequence in which courses were taken; however, it may be the case that the sequence occurs over multiple semesters and the patterns do not necessarily represent courses taken in sequential semesters.

Results for sequential rule mining. Although Sequential Pattern Mining does provide important information on the most frequent course-taking subsequences and the number of students whose academic profiles contain such subsequences, there is no assessment of the probability that a pattern will occur. To address this problem, I used sequential rule mining to discover sequential rules in students' course-taking sequences. These rules provide interesting insights into sequential patterns by giving a measure of confidence of whether a sequence of course-taking pattern would occur. For example, a rule such as (Calculus I) => (Calculus II) with a 45% confidence means that we can predict that a student taking Calculus I will later take Calculus II with a 40% confidence. The rule mining analysis returns a large number of rules alongside measures of their confidence and minimum support. Since I am specifically interested in rules that include switching or quitting out of the University, I focus only on rules that can predict when a specific course is taken, whether it is likely to be followed by switching the major or leaving the University. From Sequential Pattern Mining results, I know that there are specific courses are more frequent in quitters' or switchers' course-taking patterns. Therefore, I focus on the rules containing such courses followed by quitting or switching.

Table 4. 12

Course-taking Rules for STEM-Considering Students Who Switched Majors or Left the University

Rules	Support	Confidence (Probability)
(Calculus I, Chemistry I) => (Quit)	381	24%
(Calculus I) => (Quit)	598	20%
(Chemistry I) => (Quit)	537	21%
(Calculus I) - (Calculus II) => (Switch)	378	20%
(Calculus I) => (Switch)	627	20%
(Chemistry I) => (Switch)	414	16%
(Calculous II) => (Quit)	369	16%
(Calculus I) - (Calculus II) => (Quit)	303	15%

Table 4.12 shows the confidence and support measure of course-taking rules for switching to non-STEM or leaving the University. The results show that the probability

of dropping out of the University after taking Calculus I is 20%. The probability is even higher for students who take Chemistry I (21%). The highest probability of dropping out of the University is when a student takes both Calculus I and Chemistry I at the same semester. The probability of dropping out of the University for such students is around 24%. Also, the probability of dropping out of the University after taking Calculus I followed by Calculus II is around 16%. These results confirm that there are gatekeeper courses such as the ones mentioned above that contribute significantly to a student's decision not only to leave their potential STEM major but also to drop out of the University all together.

Interestingly, these are the same courses that might also lead to switching to a non-STEM major. The results show that a student who takes Calculus I has a 20% chance of leaving his/her field to a non-STEM major. There is the similar chance of switching to non-STEM fields after taking Chemistry I. The highest probability of switching to non-STEM, however, belongs to students who take Calculus I followed by calculus II. These students have a 19% chance of switching from a STEM to a non-STEM major. In other words, taking introductory mathematics courses is directly related to leaving STEM fields for a non-STEM major.

Student performance in gatekeeping courses. As a follow up step, I looked at student grades for selected courses to better understand how course taking patterns might contribute to students' decisions to switch to a non-STEM major or leave the University. In this analysis, I considered what appears to be three key potential gate keeping courses: Calculus I, Calculus II, and Chemistry I. As noted above, there appears to be consistent patterns for switching away from STEM majors and quitting the University that occur after taking these courses. One potential reason for this could be student performance in these classes.

Overall performance in gate keeping courses. Table 4.13 describes the percentage of STEM-considering students who received specific letter grades for Calculus I or withdrew from the class prior to receiving a grade. Overall, 84% of students who took Calculus I passed the course with a grade of C or above, and about 5% withdrew. However, among STEM-considering students who ultimately graduated with a STEM-related degree, almost 90% passed the course with a letter grade of C or above and 40% receive A's. This is in contrast to students who were initially a STEM major and later switched to a non-STEM major and those that later left the University after having taken Calculus I. Among quitters, only 65% passed, just 19% received A's, and 10% failed the course and another 10% withdrew. The pattern was less clear for switchers – that said, on average, switchers received lower grades in Calculus I compared to students who graduated with STEM majors.

Table 4.14 considers student grades for Calculus II. Here we find that about 84% of STEM-considering students pass this course with a grade of C or better; however, 9% of students withdraw the course. Again, there are descriptive differences in grades among students who graduate with a STEM major, those that switch away from STEM majors, and those that leave the University. Ninety percent of students who graduate with a STEM degree, who also take Calculus II, pass the course and 6% of these students withdrew. In contrast, students who left the University were more likely to have failed or withdrew from the course (12% and 16%, respectively). Quitters who passed the course also, on average, received lower passing grades. Among switchers 82% passed the course but did so with lower average grades – for example, just 22% received A’s (compared to 37% of STEM graduates). Also, about 13% of switchers withdrew from the course (compared to 6% of STEM graduates).

Table 4.15 presents the distribution of grades for Chemistry I. Overall, 78% of STEM-considering students who took this course passed with a grade of C or better, and 9% withdrew from the course. However, among students who graduated with STEM-related degrees, 88% passed the course – although just 14% received A’s, and 6% withdrew. By contrast, among students who left the University, 58% passed the course – with just 4% of students receiving an A. Eleven percent of students who quit the University failed this course and another 16% withdrew. This suggest that nearly 150 students across three cohorts left the University after failing or withdrawing from this course. Although 71% of switchers passed the course, however, 13% of switchers withdrew before its completion. Taken together, the descriptive patterns in student grades in these three courses suggest that student performance (i.e., grades) may be a contributing factor to STEM-considering students’ decisions to pursue a STEM-related degree, as well as whether they decide to remain at the University.

Table 4. 13
Distribution of Grades for Calculus I

Calculus I	Overall			STEM Graduates			Switchers			Quitters		
Grade	total	male	female	total	male	female	total	male	female	total	male	female
A	33%	27%	40%	40%	34%	46%	27%	21%	35%	19%	16%	25%
B	31%	32%	29%	32%	34%	30%	32%	33%	25%	24%	23%	28%
C	20%	22%	17%	17%	29%	16%	24%	26%	21%	22%	26%	17%
D	8%	8%	6%	5%	5%	5%	9%	10%	7%	15%	15%	13%
F	4%	5%	3%	2%	2%	1%	3%	4%	2%	10%	11%	10%
W	5%	6%	5%	4%	4%	4%	5%	5%	5%	10%	11%	8%
Total	3,267	1,734	1,533	1,911	972	939	683	353	330	673	409	264

Table 4. 14
Distribution of Grades for Calculus II

Calculus II	Overall			STEM Graduates			Switchers			Quitters		
Grade	total	male	female	total	male	female	Total	male	female	total	male	female
A	31%	26%	39%	37%	32%	45%	22%	18%	27%	18%	14%	27%
B	33%	34%	31%	34%	37%	32%	34%	32%	35%	23%	24%	21%
C	20%	22%	18%	18%	21%	15%	24%	28%	20%	24%	23%	25%
D	4%	5%	3%	3%	3%	2%	6%	5%	6%	10%	12%	5%
F	3%	3%	2%	1%	1%	1%	2%	3%	1%	10%	12%	6%
W	9%	9%	8%	6%	7%	6%	11%	13%	8%	16%	16%	15%
Total	2,537	1,464	1,079	1,673	945	728	459	253	206	411	266	145

Table 4. 15

Distribution of Grades for Chemistry I

Chemistry I	Overall			STEM Graduates			Switchers			Quitters		
Grade	total	male	female	total	male	female	total	male	female	total	male	female
A	10%	10%	11%	14%	14%	11%	4%	3%	5%	4%	3%	6%
B	31%	31%	32%	38%	39%	37%	26%	23%	29%	16%	16%	17%
C	37%	29%	26%	36%	37%	36%	41%	44%	39%	38%	39%	35%
D	9%	10%	7%	6%	5%	5%	12%	14%	10%	15%	16%	14%
F	3%	4%	3%	1%	1%	1%	3%	3%	3%	11%	11%	11%
W	9%	8%	11%	6%	4%	8%	13%	13%	14%	16%	16%	17%
Total	2,572	1,310	1,262	1,583	786	797	424	192	232	565	332	233

Gender differences in performance. Descriptively, average grades in Calculus I were different for female and male STEM-considering students. Eighty-six percent of female students passed Calculus I with a grade of C or higher, while 72% of male students passed with similar grades. That said, nearly 40% of female STEM-considering students received an A grade, while 27% males received an A. A similar pattern was apparent for Calculus II (Table 4.14). Overall, among STEM-considering students, females were slightly more likely to pass Calculus II than their male counterparts (88% vs. 82%, female vs. male). However, for Chemistry I, female and male students were equally likely to pass the course with a grade of C or better. All that said, while there were general differences among male and female STEM-considering students in the grades received in Calculus I, Calculus II and Chemistry I, there were no discernible patterns that suggested that grades contributed to gender differences (described above) in the share of women and men who switched from a STEM-related majors to a non-STEM degrees. In Tables 4.13-4.15 we see comparable distributions in grades among men and women who were STEM graduates, switchers, and quitters.

Gender differences in dominant course taking patterns. To get a better understanding about gender differences in course-taking patterns I used discriminant subsequent analysis to investigate whether there is an association between student's gender and course taking patterns. Pearson independent Chi-square is applied to measure the strength of association of each subsequence with the covariate (gender) and then selects the subsequences with the strongest association. This analytic approach identifies the most different course taking patterns between female and male students who initially declared a STEM major. The frequencies of all 20 subsequences that significantly discriminate for the gender at (*p< 0.01*) level are plotted in Figure 4.4. The colors used for the bars in the figure indicate the sign and significance of the associated Pearson residual.

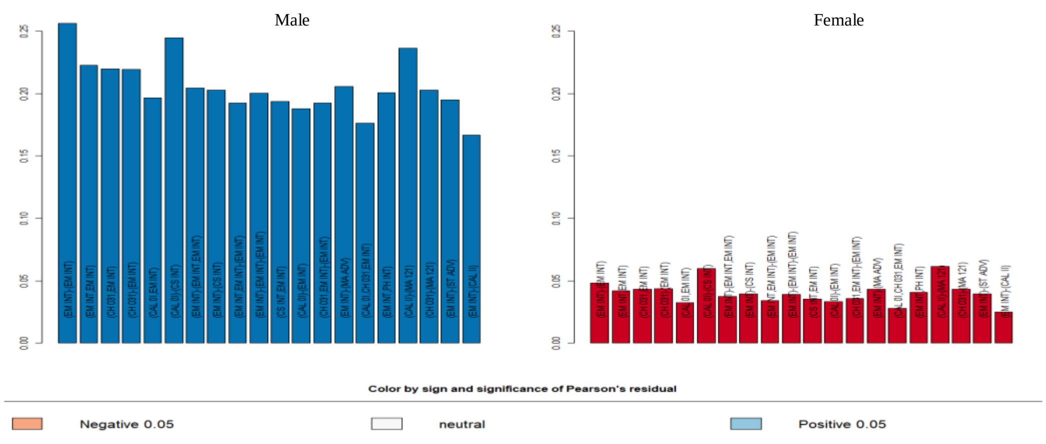

Figure 4. 4. *Course taking subsequences that discriminate gender at the 1% level*

Note: Blue: Positive 0.01 and Red: Negative 0.01

Table 4.16 presents the most discriminating course-taking subsequences in decreasing order of their discriminant power with the frequencies for male and female students. The most discriminant one is the one with the highest Chi-square. As the table shows, most of the top 20 discriminant course taking patterns include at least one engineering or computer science course. Men are also more likely to take Calculus I (CAL 0I) and Introductory Engineering (EM INT) concurrently (20% male vs. 3% female), or Calculus I (CAL 0I) and then Introductory Computer Science (CS INT) (24% male vs. 6% female). Similarly, Calculus III (MA 121) also is more likely to appear in a course taking sequence among male students than female students. For example, 24% of male students took Calculus II followed by Calculus III (MA 121), while only 6% of female students followed such course-taking pattern. This finding is consistent with the differences in academic majors between male and female students discussed above. That said, discriminant analysis only gives me the most discriminant course taking patterns, but it does not tell me if there are courses taking patterns that females are more likely to take than male students. To answer this question, in the next section I use cluster analysis to investigate whether students' course-taking pattern is divided along gender line.

Table 4. 16

Course taking subsequences that discriminate gender at the 1% level

Subsequence	Support	Chi-2	Freq. Male	Freq. Female
(EM INT) → (EM INT)	0.14	424	0.26	0.05
(EM INT, EM INT)*	0.13	360	0.22	0.04
(CH 031, EM INT)*	0.12	345	0.22	0.04
(CH 031) → (EM INT)	0.12	342	0.22	0.04
(CAL 0I, EM INT)*	0.11	337	0.20	0.03
(CAL 0I) → (CS INT)	0.15	333	0.24	0.06
(EM INT) → (EM INT, EM INT)*	0.11	332	0.20	0.04
(EM INT) → (CS INT)	0.12	318	0.20	0.04
(EM INT, EM INT)* → (EM INT)	0.11	316	0.19	0.03
(EM INT) → (EM INT) → (EM INT)	0.11	312	0.20	0.04
(CS INT, EM INT)*	0.11	312	0.19	0.04
(CAL 0I) → (EM INT)	0.10	308	0.19	0.03
(CH 031, EM INT)* → (EM INT)	0.11	306	0.19	0.04
(EM INT) → (MA ADV)	0.12	305	0.21	0.04
(CAL 0I, CH 031, EM INT)*	0.10	305	0.18	0.03
(EM INT, PH INT)*	0.11	304	0.20	0.04
(CAL II) → (MA 121)	0.14	303	0.24	0.06
(CH 031) → (MA 121)	0.12	298	0.20	0.04
(EM INT) → (ST ADV)	0.11	296	0.19	0.04
(EM INT) → (CA L II)	0.09	295	0.17	0.02

Note: A complete list of full course titles alongside their designated code appears in Appendix C.
* notes courses that are taken concurrently.

Cluster analysis. I developed a typology of students based on their course taking behavior using cluster analysis techniques. The resulting typology helps us to understand whether specific course taking behaviors are associated with gender – i.e., are certain course taking behaviors more likely for women or men.

To develop the clusters, I computed the normalized OME (Optimal Matching Event) dissimilarity matrix for all STEM-considering students. This resulted in a dendrogram plot clustering tree (see Figure 4.6). This plot identifies 12 groups of students who have similar course taking patterns. To better understand each group's course taking behavior, I looked within clusters to find the most frequent course sequence patterns. I gave each cluster a name that described the dominant course taking patterns contained in that cluster, and then I identified the share of students in each cluster that were male and female. Appendix D lists the most frequent course sequencing patterns by cluster.

Table 4.17 summarizes the 12 clusters and the distribution of male and female students within each cluster. The three sequential course taking clusters with the most STEM-considering students were: 1) Non-STEM and Switchers (14%); 2) Switchers (13%); and 3) Non-STEM (11%). Altogether, these three clusters are comprised of 38% of STEM-considering students. Two clusters describe student course taking patterns for students who left the University (i.e., Quitters 1 and 2; 10% and 5%, respectively). Four clusters identified 19% of students with course taking patterns in in life science or related subfield. Ten percent of students were identified by course sequences related to an engineering major.

Table 4.17

Student Clusters Based on Course Taking Patterns

Clusters	Total Students (%/n)	Male (%/n)	Female (%/n)	% Difference Between Male/Female
Non-STEM and Switchers	14% (646)	6% (272)	8% (374)	2% (172)
Switchers	13% (606)	6% (282)	7% (324)	1% (42)
Non-STEM	11% (553)	2% (101)	9% (452)	7% (351)
Engineering	10% (576)	8% (389)	2% (87)	6% (302)
Quitter 1	10% (523)	5% (227)	5% (251)	~ 0%
Life Science	10% (472)	4% (207)	6% (265)	2% (58)
Life science/Agriculture	7% (314)	2% (86)	5% (228)	3% (142)
Math and Computer Science	7% (328)	5% (225)	2% (103)	3% (112)
Life Science, with Chemistry and Biology	7% (316	3% (141)	4% (175)	2% (34)
Life Science/Food Science	5% (226)	1% (32)	4% (194)	3% (162)
Quitter 2	5% (249)	4% (204)	1% (45)	3% (159)
Nursing and Health Science	2% (95)	1% (33)	1% (62)	~ 0%
Total	4890	2,265	2,625	

In an ideal situation, male and female students should be equally distributed across course taking pattern clusters. But what I find is that some course taking patterns are comprised of more male or female students. Specifically, the six course taking clusters that were most dissimilar in female and male membership were: 1) Engineering; 2) Non-STEM; 3) Food Science; 4) Agriculture; 5) Math & Computer Science; and 6) Quitter 2. For example, while the engineering cluster is comprised of 10% of STEM-considering students, 8% of these students are male and just 2% are female. Conversely, in the non-STEM cluster comprised of 11% of students, 9% of these students are female and only 2% are male. Interestingly, the clusters that captured course taking patterns related to life science tended to have more females than males (e.g., food science, 4% vs. 1%, female vs. male). Finally, female students have more presence in the clusters that capture course-taking patterns leading to switching to a non-STEM fields whereas male students are more likely to belong to clusters that capture a pattern of dropping out of the University. Taken together, these findings are consistent with differences between male and female declared academic majors. That is, male students tend to be more represented in engineering course taking patterns, while women are more likely to pursue course taking patterns in life science.

To conclude, using Sequential Pattern Mining techniques, I was able to identify the most frequent patterns in STEM-considering students' course taking sequences. The findings revealed that introductory courses in mathematics and sciences, like Calculus I, Calculus II, Chemistry I, and others are among the most frequent courses taken by STEM-considering students. More specifically, my findings show that Calculus I, Calculus II, and Chemistry I are frequently present in broken course taking patterns, i.e., patterns in which a student who initially declared a STEM major later switches or drops out of the University. This leads to the conclusion that these courses might be acting as gatekeepers, discouraging students from taking more STEM courses and pushing them to move to other fields or even to drop out of the University. Furthermore, my investigation of student performance patterns in these courses leads me to believe that student performance (i.e., grades) may be a contributing factor in STEM-considering students' decision whether to switch out of a STEM-related degree or even to drop out of the University all together.

My findings also point to a strong association between gender and course-taking patterns. Taking engineering and computer science courses is a significant male course-taking behavior, for example. Clustering students' course-taking patterns allowed me to identify additional course taking patterns in which gender seems to play a significant role. Female students, for example, mostly follow course taking patterns that are heavy in life science. Women are also slightly overrepresented in clusters with course-taking patterns that lead to switching to non-STEM. Male students, in contrast, are overrepresented in course-taking clusters that lead to quitting from the University. Interestingly, I did not see a significant difference in the distribution of grades between men and women who were STEM graduates, switchers, and quitters. More extensive research is needed to understand why more women are leaving STEM fields for non-STEM fields in spite of the fact that there is no significant difference between their performance and that of their male counterparts. Having said that, what my results highlight is the fact that any narrative that tries to explain away this disparity by taking recourse in the issue of poor academic performance is simplistic and based on unwarranted assumptions.

CHAPTER 5: DISCUSSION

In spite of all the investment in the last few decades on STEM education, low enrollment and high attrition rate among students in these fields remain an unmitigated challenge for institutions of higher education. The underrepresentation of women and minority students in such fields replicates itself in the makeup of the workforce, adding another layer to the challenge. Although previous research has provided valuable information about enrollment and attrition rates and insightful analysis of some of the factors contributing to such patterns, there are still many questions that remain unanswered. We know that a student's decision to declare a major, stay in one, or leave it is influenced by choices made at different points of his/her college career under different circumstances. One important factor is how the student interacts with the curriculum and his/her experience of such interactions. The college curriculum in any given major is an academic plan developed and structured by faculty, program directors, and the administration with the goal of enhancing students' learning and achieving a certain level of literacy in a given field. Whether such plans are successful in reaching their goals depends, in part, on how the student experiences them. The experience of interaction with the curriculum is a complex and multilayered one influenced by different components of the curriculum including content, pedagogy, and instructional resources, the faculty, and other external factors (Cohen & Kisker, 2012). Given the complexity of this experience and variety of factors influencing it, it is impossible to fully capture this experience. There have been efforts, however, to capture some aspects of this experience using tools such as course evaluations, surveys, enrollment history, etc. Detailed student transcripts are an important piece of multidimensional data that can be used for this purpose. The transcript is like a history map of the student's academic progress. If properly analyzed, it can provide us with valuable insights into the student's experience in navigating the curriculum and interacting with it and how it influences in her/his decision-making process in following different academic paths overtime.

The methodological approach used in this study was a first attempt to apply data mining methods to use rich multidimensional data to enhance our understanding of student academic behavior/paths and determine identifiable patterns that emerge from the actual course-taking experiences of the students as they progress through their study. The identified patterns help us understand the ways in which the college's curriculum might help or hinder student progress in STEM fields and how it can favor student groups who already dominate such field, leading to further marginalization of underrepresented groups such as women and/or racial/ethnic minorities. In the rest of this chapter, I will synthesize the study's findings. This is followed by a discussion the study's implication.

The Unpopularity of STEM Trajectories

The descriptive statistics presented in this study reaffirms that STEM majors are much less popular than non-STEM majors. Despite the allure and promise of economic success for STEM graduates, most of the students in this sample chose to not pursue degrees on STEM fields. Among the minority of students who chose to pursue degrees in STEM fields, life science was the most popular field. By contrast, a comparatively smaller share of students pursued a major in other STEM fields such as engineering or computer science. Low enrollment patterns in STEM fields in general, and in hard sciences in particular, are consistent with findings from other national studies (Chen, 2013; Pryor et al., 2010; Snyder & Dillow, 2011).

Not only STEM fields are unpopular, as discussed above, at the point of entry to college, but also, trajectory analysis reveals, this unpopularity increases as students go further in their studies. That is to say, even for the students who enter college with the intention to study in STEM fields, many of them later change their minds and switch to a non-STEM field. A similar pattern has been confirmed in previous studies that have examined the problem of student persistence in STEM (Chen, 2013, 2015; Kokkelenberg & Sinha, 2010). All these studies clearly show that more students switch from STEM to non-STEM compared to the other way around (e.g., Chen, 2015; Griffith, 2010). None of these studies, however, were able to look at student trajectories over time, and/or follow the students' major status for each point of their study throughout their college career until they graduate or drop out. Another shortcoming of the previous studies has been their reliance on reported majors (mostly self-reports), which makes their findings less accurate and reliable. My approach offers a remedy for this shortcoming. My detailed analysis of transition patterns shows that 19% of the students switch from STEM fields to non-STEM fields compared to only 1.5% of non-STEMs switching to a STEM field.

Finally, my study also reveals a significant difference in student majoring trajectories within STEM paths. As stated before, compared to life sciences, hard science majors (such as engineering and computer science) are much less popular among students at the point of entry. In addition, their unpopularity grows as students go further in their studies. Although student attrition of any major/department can be a cause for concern for certain constituencies, the results of this study point out to a particularly serious area of national concern given the fact that majors such as engineering and computer science have been identified by the federal government as areas of needs. Attracting enough students and keeping them in these fields are essential for the development of a skilled workforce in the national level that can help the nation compete in the 21st century global economy

Female vs. Male Major Trajectory Differences

Another important finding of this study is the detailed patterns of difference that it reveals between female and male trajectories in STEM. The results show that female and male students follow clearly different majoring paths in general, and within STEM fields in particular. More females follow non-STEM paths and their numbers even grow larger as they continue to make progress in their studies. Within STEM fields, life science trajectories have much higher popularity in general, as discussed above. This pattern seems to be driven largely by female students, who have a clear preference for life sciences over hard sciences such as engineering and computer science. For male students, on the other hand, soft and hard science trajectories seem to be distributed more equally. In spite of the clear patterns discussed, since most of the previous studies on the subject have been unable to track students' majoring patterns overtime, some have expressed doubts about whether a meaningful gender disparity exists in terms of persistence in STEM fields. This study's unique methodology, however, allows us to do disparity analysis for both STEM in general and for specific STEM fields in particular. The results suggest that, at least for the institution represented in this study, there is a gender disparity when it comes to student interest in life sciences in comparison to hard sciences. The results show that women leave hard sciences for non-STEM majors with a higher rate compared to men. In addition, even those women who stay in STEM fields switch to life sciences in higher numbers compared to their male counterparts. This is a clear testimony to the fact that the institution under study has failed to recruit and keep women in STEM and particularly in hard sciences even though job prospects and the prospect of receiving a better compensation package are higher in such fields compare to non-STEM fields (Xie, Fang, & Shauman, 2015). As Hill and colleagues (2010) argue, if we lived in an ideal society in which bias and stereotypes against women in such fields did not exist, we could interpret these results merely as a reflection of females finding their passion in fields other than males. The society in which we live, however, is far from ideal and we know that women's hesitation to enter and staying in these fields is influenced negatively by cultural, social, and institutional factors that create major hurdles in the path of women who would potentially be highly successful in such fields (Fox et al., 2009). In such a situation, the least our institutions can do is, after acknowledging the problem, invest time, energy, and financial resources to find creative ways to decrease the impact of such overarching factors to the extent possible.

Gatekeeper Courses

To better understand how student course-taking patterns contribute to their decision to leave or stay in STEM fields, I conducted sequential pattern analysis in my study. This analysis helps to identify course taking patterns that are most frequently followed by STEM students as well as the students who switch to other fields or eventually leave the college. The results show that specific introductory courses in mathematics, that is, Calculus I and Calculus, II, as well as Chemistry I, play a critical role in overall student persistence in STEM fields in this institution. The specific courses mentioned above also appear repeatedly in course taking patterns of students who choose to leave STEM fields for other majors or to drop out of college. These findings suggest that these courses might be acting like a gatekeeper for STEM, blocking many students from making progress in their pursuit of their chosen STEM major and pushing them to transition to other fields or even drop out of college.

The results of my analysis align with previous studies that focus on the relationship between taking introductory STEM courses and the dropout or the switching rate (e.g., Chen, 2013, 2015; George-Jackson, 2011; Griffith, 2010). These studies, however, were unable to identify the exact culprit courses. Rather, their analysis is based on some initial speculation and conjecture on the part of the authors. An assumption is made, for example, that math introductory courses are probably hindering students from going further in their course of study. Based on this assumption, the authors focus on the relationship between taking introductory math courses and the student dropout or switch rates in their analysis of the data. In other words, whereas in my methodology the relationship between specific courses and the dropout or switch rate directly emerges from the data set analysis, previous studies have had to make assumptions about possible relationships between a category of courses (introductory math) and the student dropout or switch rate.

To understand and examine student course-taking patterns in more depth, I have used clustering, which allows for similar course-taking patterns to emerge, helping us to identify distinct groups. Analyzing these groups' course-taking patterns provides important insights into their behavior. For example, my analysis reveals that students who switch to non-STEM fields have similar course-taking patterns and can be thus clustered in a group called "the Switchers." If we look at this group's frequent course-taking patterns, we find that Calculus I followed by Calculus II is the most frequent pattern of course taking followed by its members. In contrast, Chemistry I and other introductory science courses have a less significant presence in this group's course-taking patterns. Another group that clustering allows us to identify based on their similar course taking patterns is "the Quitters," consisting of students who eventually drop out of college after having declared a STEM major. Among the members of this group, Calculus I and Chemistry I, taken together, is among the most common course-taking patterns. The identification of such specific patterns allows us to focus our attention on specific courses that seem to be co-related with STEM attrition and possibly launch more in-depth studies in order to find what exactly is causing the problem.

Another key dimension of emerging clusters that need to be discussed is how gender dynamics interact with them. Looking at the Switcher group for example, my results are clear that the number of women associated with this group is slightly more than men. In contrast, male students had more presence in the Quitter group, which consists of students who drop out of college. More importantly, my analysis of student performance levels reveals an important gap between the two genders and the possible reasons why they leave STEM for other fields. My findings show that female students' performance in a number of introductory STEM courses is meaningfully different than that of male students.

For example, focusing on the Switchers, a large number of male members of this group withdrew or failed in Calculus II while a considerable number of women Switchers attained grade A in this course. This important difference leads us towards a preliminary conclusion: That women's decision to leave STEM might have less to do with them finding the courses hard and their lower than expected academic performance and more to do with other factors that are related to broader cultural issues rather than the course content. These results align with other studies in the field that demonstrate that high achiever students, especially women, are still prone to leave STEM fields for other majors (Brainard & Carlin, 1998; Chen, 2015; Lowell et al., 2009).

Shifting focus to the issue of performance among the Quitters, my results clearly show that student performance in Chemistry I, which is a course that appears frequently in the Quitters course-taking patterns, does not follow the same pattern. Most students in this group, irrespective of their gender, performed poorly in this course. This adds an additional layer of nuance to our discussions. The results overall suggest that there are specific courses that have a significant function in hindering or blocking student progress in STEM trajectories. Having said that, it appears that not all these courses function in the same way. Rather, they might contribute to the student decision not to pursue a STEM major, or a college degree, in different ways.

Finally, the associate rule mining analysis provided even more insight into the predictability of a student's decision to leave the fields after taking the gatekeeper courses. Taking Chemistry I and Calculus I together, for example, has the most predictability power when it comes to STEM students making the decision to leave the University without earning a degree (24%). The probability of leaving the fields for a non-STEM field after taking Calculus I is around 20%. Even when a student decides to stay in this field after taking Calculus I, there is still the same chance of switching to another field after taking Calculus II. Needless to say, these are high probabilities. If we want to improve student retention rates in STEM fields, the results are really useful in pinpointing exactly where the problem needs to be tackled for the most effective results.

Implications

Major implications of this study can be classified and discussed under two broad categories: conceptual and empirical. Conceptually, the study provides us with a new framework for considering student academic trajectories in STEM fields, and empirically, the study contributes to existing knowledge about student course taking patterns and academic major selection, and potential differences between male and female students.

Conceptual implications. Most studies examining student pathways in STEM pipelines have focused on student outcomes and their relationship to some individual and institutional factors and have not described students' academic experiences or progress. In short, existing studies provide a snapshot of student experience, but tell us little how students interact with the college curriculum as they progress toward degree. In fact, past research has repeatedly suggested that more work is needed to understand student curricular experience throughout the STEM pipeline and have called for developing new methodologies to enable researchers to focus on the process rather than outcome and investigate how it influences student decision-making (Bahr, 2013; Chen; 2013; Shapiro & Sax, 2009).

In contrast, this study conceptualizes student experiences differently. Specifically, rather than considering student experiences as a collection of courses taken, this study reconceptualizes student experience as a process, with sequential pathways through the curriculum toward a degree. To do so, the study leverages detailed transcript data to deepen our understanding of student curricular experiences seen as process as they make progress in their studies.

In doing so, the study conceptualizes student experiences as a multidimensional process, evidenced in the sequential structure of transcript data. This allows us to describe the dynamic process evidenced in how students proceed toward degree – both in terms of the sequence of courses taken and also the ebb and flow of when and which students identify academic majors. Most prior studies fail to conceptualize, or describe, the sequential nature of a student's academic experience and none describe the subsequent variations in majoring or course-taking patterns that manifest themselves as the students make progress along their pathway. For example, previous studies that have examined major transition in STEM pipeline have been unable, due to the limitations of their methodology, to track student majoring trajectories at each point of time throughout their college career, relying only on major reports gathered in few data points in their study. The results are unreliable and inaccurate because many changes could happen between those data points, a shortcoming that my study remedies.

Second, this study broadly conceptualizes student course taking in STEM, allowing us to examine how students move among the full complement of STEM-related courses and majors offered at the University. While a number of studies have been conducted with the aim of investigating patterns in order to discover how student curricular experience influences the STEM pipeline (e.g., Bahr, 2013; Wang, 2016), most considered only one or two subjects at a time. In contrast, my study examines a more complete set of courses taken by the totality of students during their whole college career. While caution must be exercised against generalizing the results that were achieved based on a case study, the analytic approach I have offered here provides researchers with a universal tool that enables them to thoroughly examine the impact of a designed curriculum on student decision-making patterns, whether to stay in the path he/she began at the point of entry or leave the field, or even leave the college without earning any degree.

Complementing the re-conceptualization of student experiences, the study adopts an innovative analytic approach to describing students' sequential pathways to degree completion. The new method, which is a data mining technique, has a number of advantages for studying student experiences with the curriculum. It requires minimal assumptions about student behavior and decision-making process as they interact with the curriculum, allowing for a more comprehensive picture of student pathways – and, one that maps more closely to actual student experiences. In contrast, most previous research examining STEM pipelines have used methods that, in essence, assume linearity and uniformity of student behavior (Bahr, 2013). The assumption of linearity, for example, is at work when some of these studies draw a relationship between student persistence outcomes and some individual and institutional characteristics. Recognizing the complexity of student experience, however, in this study I have been able to identify actual patterns that emerge out of each student's complete course-taking patterns without imposing such assumptions.

More importantly, the new method introduced here has a unique visualization tool that provides us with a visual representation of the entire range of student trajectories. To quote Tukey, "The greatest value of pictures is when it forces us to notice things that we never expect to see". The visual plots created by this tool help us to interpret the resulting patterns and students majoring trajectory changes in different points of time, compare male and female trajectories, and link them to their decisions (such as switching from STEM to non-STEM or within STEM) and outcomes. Visual representation can also be extremely useful in facilitating better understanding and communication among the faculty, program directors, and other stakeholders on how STEM students' trajectories are changing over time and how different group of students are following different paths.

Empirical implications. Based on the most frequent course-taking patterns identified for students who switch to a non-STEM major, we can identify introductory math courses such as Calculus I and II as gatekeepers that impede student progress in STEM. Additionally, we know that some of these leavers, especially women, perform academically well in these courses. Taken together, these results suggest that particular elements of the STEM curriculum, especially some introductory courses, are discouraging students from continuing their study in such field. The implication being that STEM programs need to evaluate their curriculum, especially their introductory courses, to find out what elements and conditions are to blame for the students' decision to leave the fields.

My study also identifies course-taking pattern that lead to drop out for this institution. For example, in the case of the college understudy, taking Chemistry I and Calculus I together is the strongest predictor of the probability of dropping out of college. When it comes to Chemistry I, additional analysis reveals that most students who later drop out have had a poor performance in this course. More specifically, when the course is offered, it usually trims around 20% of the students. A usual interpretation of this trimming rate is that only 80% of the enrolled students on average are prepared to continue their study in STEM. An alternative way of looking at the rather high trimming rate of the course might be that a good number of students who intend to enroll in this course need extra preparation beforehand. In sum, my study's findings show that there are clear challenges when it comes to student experience with some specific introductory courses in STEM. Based on the observed patterns, universities may wish to reevaluate their curricula, particularly introductory course offerings.

Finally, this study demonstrates how data mining can be used to consider students' academic progress and explore how their students are navigating and interacting with curriculum. It can highlight and identify the exact phases of study in which the students are more at risk of leaving the programs, allowing for the people who are involved in running the program to make timely and effective interventions to prevent attrition. It can also help academic advisors to identify course-taking patterns that put specific groups of students at the risk of dropping out or switching to other fields. They can advise the students to take some particularly challenging courses in a specific sequence or to take preparatory introductory courses beforehand.

Limitation of the Study and Recommendations for Future Research

Students' course-taking patterns are influenced by different factors such as their relationship with particular instructors, the amount of course work, time table, the classroom climate, etc. An ideal analysis model would take all these factors into account. Due to the limitations of available data, however, this study is not able to include all the relevant factors. The data provided by the University, for example, does not include a complete list of course instructors. Even when we do have access to the relevant data, including some of them requires additional or different tools of analysis that are not available yet. For example, to include information about faculty in my analysis requires the deployment of a hierarchical sequential pattern mining, which is still under development. As another example, including available information about student cohorts who take a specific course together requires an additional layer of social network analysis which is beyond the scope of this study.

Although the data mining techniques employed here provide us with valuable information regarding student academic pathways and their course-taking pattern, it is important to note that these patterns or rules are not causal in nature. In other words, we cannot draw a causal inference from patterns or rules provided by data mining techniques here. Additional empirical research needs to be conducted for such causal relationships to be established between student course taking patterns and their academic outcomes.

The method proposed here was applied only to data from the institution with limited racial and ethnic diversity in its student body. Future research could apply this method to a large and more diverse institution to find out how course-taking patterns vary among students coming from different ethnic and/or racial background. From previous research, we know that Black and Hispanic students have very low enrollment and high attrition rates in STEM fields. This method could be applied to find out whether there is race-based course-taking patterns that increase the probability of leaving the fields or dropping out of college. In addition, my data did not provide information about students' financial aid situation or their socioeconomic status, which previous research have strongly associated with attrition rate in these fields. Future research could investigate course-taking patterns for students with different SES or financial aid status to better understand their academic behavior.

Another limitation of this study was that I did not have access to previous performance records of the students (like high school or middle school performance records). We know from prior studies that such records are important for the study of student performance levels at college. Students who have been academically less prepared in their K-12 years are more likely to switch out of STEM or leave college all together. Future research could incorporate the study of student academic preparation prior to their arrival at college and explore whether more prepared students take different paths to obtaining a degree in STEM or not and how this preparedness affects their outcome. One can also explore whether certain course-taking patterns are more likely to block less prepared students from pursuing a degree in STEM fields. The results could be used by advisors to help less prepared students to take courses in a pattern that could potentially help them to progress more successfully in their course of study.

Finally, since this study was conducted at one university, its findings cannot be generalized. It would be interesting to apply this method to data from a nationally representative transcript data. Majoring and course-taking patterns identified by such data will allow us to draw more general conclusions about STEM fields and students' course-taking patterns in these fields at the national level. For example, one of the gatekeeper courses found in this study is Chemistry I. A natural question is whether this course functions in this particular way only at this particular institution or is it symptomatic of a larger pattern throughout many other colleges and universities.

REFERENCES

Adelman, C. (1999). *Answers in the toolbox: Academic intensity, attendance patterns, and bachelor's degree attainment*. Washington, DC: U.S. Department of Education.

Adelman, C. (2004). Principal indicators of student academic histories in postsecondary education, 1972-2000. U.S. Department of Education.

Adelman, C. (2005). *Moving into town--and moving on: The community college in the lives of traditional-age students*. U.S. Department of Education.

Adelman, C. (2006). *The toolbox revisited: Paths to degree completion from high school through college.* U.S. Department of Education.

Aggarwal, C. C. (Ed.). (2007). Data streams: Models and algorithms. *Springer Science & Business Media, 31*.

Agrawal, R., & Srikant, R. (1995, March). Mining sequential patterns. In *icde*, *95*, 3-14.

Bahr, P. R. (2010). The bird's eye view of community colleges: A behavioral typology of first-time students based on cluster analytic classification. *Research in Higher Education*, *51*(8), 724–749. https://doi.org/10.1007/s11162-010-9180-5

Bahr, P. R. (2013a). The aftermath of remedial math: Investigating the low rate of certificate completion among remedial math students. *Research in Higher Education*, *54*(2), 171–200. https://doi.org/10.1007/s11162-012-9281-4

Bahr, P. R. (2013b). The deconstructive approach to understanding community college Students' pathways and outcomes. *Community College Review*, *41*(2), 137–153. https://doi.org/10.1177/0091552113486341

Bebe-vroman, M., Juniewicz, I., Lucarelli, B., Fox, N., Nguyen, T., & Tjang, A. (2017, March 8-11). Exploring gender diversity in CS at a large public R1 research university. *SIGCSE*, Seattle, WA. https://dl.acm.org/citation.cfm?doid=3017680.3017773

Beyer, S. (2014). Why are women underrepresented in Computer Science? Gender differences in stereotypes, self-efficacy, values, and interests and predictors of future CS course-taking and grades. *Computer Science Education*, *24*(2–3), 153–192. https://doi.org/10.1080/08993408.2014.963363

Bowen, W. G., Chingos, M. M., & McPherson, M. S. (n.d.). Completing college at America's public universities.

Brainard, S. G., & Carlin, L. (1998). A six-year longitudinal study of undergraduate women in engineering and science. *Journal of Engineering Education*, *87*(4), 369-375.

Burger, C., Abbott, G., Tobias, S., Koch, J., Vogt, C., & Sosa T. (2007). Gender equity in science, engineering, and technology. S. Klein, (Ed.), *Handbook for Achieving Gender Equity through Education, 2* (pp. 255-279).

Ceci, S. J., & Williams, W. M. (2010). Sex differences in math-intensive fields. *Current Directions in Psychological Science*, *19*(5), 275–279. https://doi.org/10.1177/0963721410383241

Center for Institutional Data Exchange and Analysis. (2000). *1999–2000 SMET retention report*. Norman, OK: University of Oklahoma.

Chang, M. J., Cerna, O., Han, J., & Sáenz, V. (2008). The contradictory roles of institutional status in retaining underrepresented minorities in biomedical and behavioral science majors. The Review of Higher Education, 31(4), 433–464.

Chang, M. J., Sharkness, J., Hurtado, S., & Newman, C. B. (2014). What matters in college for retaining aspiring scientists and engineers from underrepresented racial groups: Retaining aspiring scientists. *Journal of Research in Science Teaching*, *51*(5), 555–580. https://doi.org/10.1002/tea.21146

Chen, X. (2015). STEM attrition among high-performing college students: Scope and potential causes. *Journal of Technology and Science Education*, *5*(1). https://doi.org/10.3926/jotse.136

Chen, X. (2013). STEM attrition: College students' paths into and out of STEM Fields. Statistical Analysis Report. NCES 2014-001. *National Center for Education Statistics*. http://ies.ed.gov/pubsearch/pubsinfo.asp?pubid=2014001rev

Chen, X. (2009). Students who study science, Technology, Engineering, and Mathematics (STEM) in postsecondary education. Stats in brief. NCES 2009-161. *National Center for Education Statistics.*

Clark Blickenstaff, J. (2005). Women and science careers: Leaky pipeline or gender filter? *Gender and Education*, *17*(4), 369–386. https://doi.org/10.1080/09540250500145072

Crisp, G., Nora, A., & Taggart, A. (2009). Student characteristics, pre-college, college, and environmental factors as predictors of majoring in and earning a STEM degree: An analysis of students attending a Hispanic serving institution. *American Educational Research Journal*, *46*(4), 924–942. https://doi.org/10.3102/0002831209349460

Cohen, A. M., & Kisker, C. B. (2010). *The shaping of American higher education: Emergence and growth of the contemporary system*. John Wiley & Sons.

Crosta, P. M. (2014). Intensity and attachment: How the chaotic enrollment patterns of community college students relate to educational outcomes. *Community College Review; Raleigh*, *42*(2), 118–142.

Davis, C. S. (1996). *The equity education. fostering the advancement of women in the Sciences, Mathematics, and Engineering*. San Francisco, CA: Jossey-Bass Inc.

Ewert, S. (2010). Male and female pathways through four-year colleges: Disruption and sex stratification in higher education. *American Educational Research Journal*, *47*(4), 744–773. https://doi.org/10.3102/0002831210374351

Felder, R. M., Felder, G. N., Mauney, M., Hamrin Jr., C. E., & Dietz, E. J. (1995). A longitudinal study of engineering student performance and retention. III. Gender differences in student performance and attitudes. *Journal of Engineering Education*, *84*(2), 151-163. https://doi.org/10.1002/j.2168-9830.1995.tb00162.x

Fournier-Viger, P., Gueniche, T., & Tseng, V. S. (2012). Using partially-ordered sequential rules to generate more accurate sequence prediction. In S. Zhou, S. Zhang, & G. Karypis (Eds.), *Advanced Data Mining and Applications, 7713,* 431–442. Berlin, Heidelberg, Germany: Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-35527-1_36

Fournier-Viger, P., Lin, J. C. W., Kiran, R. U., Koh, Y. S., & Thomas, R. (2017). A survey of sequential pattern mining. *Data Science and Pattern Recognition, 1*(1), 54-77.

Fox, M. F., Sonnert, G., & Nikiforova, I. (2009). Successful programs for undergraduate women in science and engineering: Adapting versus adopting the institutional environment. *Research in Higher Education, 50*(4), 333–353. https://doi.org/10.1007/s11162-009-9120-4

Friedkin, N. E., & Thomas, S. L. (1997). Social positions in schooling. *Sociology of Education*, *70*(4), 239. https://doi.org/10.2307/2673266

Gabadinho, A., Ritschard, G., Studer, M., & Müller, N. S. (2009). *Mining sequence data in R with the TraMineR package: A user's guide*. Technical report, Department of Econometrics and Laboratory of Demography, University of Geneva, Geneva. http://mephisto. unige. ch/TraMineR.

Gabadinho, A., Ritschard, G., Müller, N. S., & Studer, M. (2011). Analyzing and visualizing state sequences in *R* with TraMineR. *Journal of Statistical Software, 40*(4). https://doi.org/10.18637/jss.v040.i04

George-Jackson, C. E. (2011). STEM switching: Examining departures of undergraduate women in STEM fields. *Journal of Women and Minorities in Science and Engineering*, *17*(2), 149–171. https://doi.org/10.1615/JWomenMinorScienEng.2011002912

Griffith, A. L. (2010). Persistence of women and minorities in STEM field majors: Is it the school that matters? *Economics of Education Review*, *29*(6), 911–922. https://doi.org/10.1016/j.econedurev.2010.06.010

Han, J., Pei, J., & Kamber, M. (2011). *Data mining: Concepts and techniques*. Elsevier.

Heck, R. H., Price, C. L., & Thomas, S. L. (2004). Tracks as emergent structures: A network analysis of student differentiation in a high school. *American Journal of Education*, *110*(4), 321–353. https://doi.org/10.1086/422789

Hill, C., Corbett, C., & St. Rose, A. (2010). *Why so few? Women in science, technology, engineering, and mathematics*. Washington, DC: AAUW.

Huang, G., Taddese, N., & Walter, E. (2000). Entry and persistence of women and minorities in college science and engineering education. *Education Statistics Quarterly*, *2*(3), 59–60.

Hurtado, S., Han, J. C., Sáenz, V. B., Espinosa, L. L., Cabrera, N. L., & Cerna, O. S. (2007). Predicting transition and adjustment to college: Biomedical and behavioral science aspirants' and minority students' first year of college. *Research*. https://doi.org/10.1007/s11162-007-9051-x

Hyde, J. S., Lindberg, S. M., Linn, M. C., Ellis, A. B., & Williams, C. C. (2008). Gender similarities characterize math performance. *Science*, *321*(5888), 494-495.

Kantardzic, M. (2011). *Data mining: Concepts, models, methods, and algorithms*. John Wiley & Sons.

Kokkelenberg, E. C., & Sinha, E. (2010). Who succeeds in STEM studies? An analysis of Binghamton University undergraduate students. *Economics of Education Review*, *29*(6), 935–946. https://doi.org/10.1016/j.econedurev.2010.06.016

Luan, J. (2002). Data mining and its applications in higher education. *New Directions for Institutional Research*, *2002*(113), 17-36.

Lowell, B. L., Salzman, H., & Bernstein, H. (2009). Steady as she goes? Three generations of students through the science and engineering pipeline. https://doi.org/doi:10.7282/T31R6S4K

Lynn, R., & Irwing, P. (2004). Sex differences on the progressive matrices: A meta-analysis. *Intelligence*, *32*(5), 481–498. https://doi.org/10.1016/j.intell.2004.06.008

Marx, D. M., & Roman, J. S. (2002). Female role models: Protecting women's math test performance. *Personality and Social Psychology Bulletin*, *28*(9), 1183-1193. https://doi.org/10.1177/01461672022812004

Mendez, G., Ochoa, X., Chiluiza, K., & de Wever, B. (2014). Curricular design analysis: A data-driven perspective. *Journal of Learning Analytics*, *1*(3), 84–119. https://doi.org/10.18608/jla.2014.13.6

Mooney, C. H., & Roddick, J. F. (2013). Sequential pattern mining -- approaches and algorithms. *ACM Computing Surveys*, *45*(2), 1–39. https://doi.org/10.1145/2431211.2431218

National Science Board. (2016). *Science and Engineering Indicators Digest*. Arlington, VA: National Science Foundation (NSB-2016-2).

Neumann, M. D., Lathem, S. A., & Fitzgerald-Riker, M. (2016). Resisting cultural expectations: Women remaining as civil and environment engineering majors. *Journal of Women and Minorities in Science and Engineering*, *22*(2), 139–158. https://doi.org/10.1615/JWomenMinorScienEng.2016013949

Newman, C. B. (2016). An inspirational and onerous journey from the great migration to the academy. In B. L. McGowan, R. T. Palmer, J. L. Wood, & D. F. Hibbler (Eds.), *Black Men in the Academy* (pp. 127–138). New York, NY: Palgrave Macmillan US. https://doi.org/10.1057/9781137567284_9

Ost, B. (2010). The role of peers and grades in determining major persistence in the sciences. *Economics of Education Review*, *29*(6), 923–934. https://doi.org/10.1016/j.econedurev.2010.06.011

Pinel, E. C., Warner, L. R., & Chua, P.-P. (2005). Getting there is only half the battle: Stigma consciousness and maintaining diversity in higher education. *Journal of Social Issues*, *61*(3), 481–506. https://doi.org/10.1111/j.1540-4560.2005.00417.x

Pryor, J. H., Eagan, K., Palucki Blake, L., Hurtado, S., Berdan, J., & Case, M. H. (2012). *The American freshman: National norms fall 2012*. Los Angeles, CA: Higher Education Research Institute, UCLA.

Rask, K. (2010). Attrition in STEM fields at a liberal arts college: The importance of grades and pre-collegiate preferences. *Economics of Education Review, 29*(6), 892–900.

Riegle-Crumb, C., Moore, C., & Ramos-Wada, A. (2011). Who wants to have a career in science or math? Exploring adolescents' future aspirations by gender and race/ethnicity. *Science Education*, *95*(3), 458–476. https://doi.org/10.1002/sce.20431

Ritschard, G. (2018). *Sequence analysis and related approaches: innovative methods and applications*. New York, NY: Springer Berlin Heidelberg.

Ritschard, G., Gabadinho, A., Mueller, N. S., & Studer, M. (2008). Mining event histories: A social science perspective. *International Journal of Data Mining, Modelling and Management, 1*(1), 68-90. 10.1504/IJDMMM.2008.022538

Ritschard, G., Gabadinho, A., Studer, M., & Müller, N. S. (2009). Converting between various sequence representations. In Z. W. Ras & A. Dardzinska (Eds.), *Advances in Data Management* (Vol. 223, pp. 155–175). Berlin, Heidelberg: Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-02190-9_8

Seymour, E. (1995). The loss of women from science, mathematics, and engineering undergraduate majors: An explanatory account. *Science Education*, *79*(4), 437-473.

Seymour, E., & Hewitt, N. (1997). *Talking about leaving: Why undergraduates leave the sciences*. Boulder, CO: Westview.

Seymour, E. (2002). Tracking the processes of change in U.S. undergraduate education in science, mathematics, engineering, and technology. *Science Education*, *86*(1), 79–105. https://doi.org/10.1002/sce.1044

Shapiro, C. A., & Sax, L. J. (2011). Major selection and persistence for women in STEM. *New Directions for Institutional Research*, *2011*(152), 5–18. https://doi.org/10.1002/ir.404

Simpson, J. C. (2001). Segregated by subject: Racial differences in the factors influencing academic major between European Americans, Asian Americans, and African, Hispanic, and Native Americans. *The Journal of Higher Education*, *72*(1), 63. https://doi.org/10.2307/2649134

Slimani, T., & Lazzez, A. (2013). Sequential mining: Patterns and algorithms analysis. *arXiv preprint arXiv:1311.0350*.

Snyder, T.D., & Dillow, S.A. (2011). *Digest of education statistics, 2010 (NCES 2011-015)*. National Center for Education Statistics, Institute of Education Sciences, U.S. Department of Education. Washington, DC.

Sprecher, S., Brooks, J. E., & Avogo, W. (2013). Self-esteem among young adults: Differences and similarities based on gender, race, and cohort (1990–2012). *Sex Roles, 69*(5-6), 264-275. https://doi.org/10.1007/s11199-013-0295-y

Steele, J., James, J. B., & Barnett, R. C. (2002). Learning in a man's world: Examining the perceptions of undergraduate women in male-dominated academic areas. *Psychology of Women Quarterly*, *26*(1), 46–50. https://doi.org/10.1111/1471-6402.00042

STEM-Education-in-the-US-2017.pdf. (n.d.). Retrieved from https://www.act.org/content/dam/act/unsecured/documents/STEM/2017/STEM-Education-in-the-US-2017.pdf

Tyson, W., Lee, R., Borman, K. M., & Hanson, M. A. (2007). Science, Technology, Engineering, and Mathematics (STEM) pathways: High school science and math coursework and postsecondary degree attainment. *Journal of Education for Students Placed at Risk (JESPAR)*, *12*(3), 243–270. https://doi.org/10.1080/10824660701601266

Voyer, D., & Voyer, S. D. (2014). Gender differences in scholastic achievement: A meta-analysis. *Psychological bulletin*, *140*(4), 1174.

Walton, G. M., Logel, C., Peach, J. M., Spencer, S. J., & Zanna, M. P. (2015). Two brief interventions to mitigate a "chilly climate" transform women's experience, relationships, and achievement in engineering. *Journal of Educational Psychology*, *107*(2), 468–485. https://doi.org/10.1037/a0037461

Wang, M.-T., & Degol, J. L. (2017). Gender gap in Science, Technology, Engineering, and Mathematics (STEM): Current knowledge, implications for practice, policy, and future directions. *Educational Psychology Review*, *29*(1), 119–140. https://doi.org/10.1007/s10648-015-9355-x

Wang, X. (2016). Course-taking patterns of community college students beginning in STEM: Using data mining techniques to reveal viable STEM transfer pathways. *Research in Higher Education*, *57*(5), 544–569. https://doi.org/10.1007/s11162-015-9397-4

Weinburgh, M. (1995). Gender differences in student attitudes toward science: A meta-analysis of the literature from 1970 to 1991. *Journal of Research in science Teaching*, *32*(4), 387-398.

Witteveen, D., & Attewell, P. (2017). The college completion puzzle: A hidden Markov model approach. *Research in Higher Education*, *58*(4), 449–467. https://doi.org/10.1007/s11162-016-9430-2

Xie, Y., Fang, M., & Shauman, K. (2015). STEM education. *Annual Review of Sociology*, *41*(1), 331–357. https://doi.org/10.1146/annurev-soc-071312-145659

Xie, Y., & Shauman, K. A. (2003). *Women in science*. Harvard University Press.

Zeidenberg, M., & Scott, M. (2011). The context of their coursework: Understanding course-taking patterns at community colleges by clustering student transcripts. CCRC Working Paper No. 35. *Community College Research Center, Columbia University*.

Zhao, Q., & Bhowmick, S. S. (2003). Sequential pattern mining: A survey. *ITechnical Report CAIS Nayang Technological University Singapore*, *1*, 26.

APPENDIX A

List of UVM's STEM Majors Based on NCES Definition of STEM fields

Description	Major
Computer Science	Computer Science
Complex Systems & Data Science	Computer Science
Computer Sci & Info Systems	Computer Science
Clinical & Translational Sci	Computer Science
Data Science	Computer Science
Bioengineering	Engineering
Biomedical Engineering	Engineering
Civil Engineering	Engineering
CE Certificate	Engineering
Civil & Environmental Engr	Engineering
Engr - Bioengineering	Engineering
Engineering Physistry	Engineering
Electrical Engineering	Engineering
Environmental Engineering	Engineering
Engineering - General	Engineering
Engineering Management	Engineering
Engineering	Engineering
Engineering Physics	Engineering
Mechanical Engineering	Engineering
Medical Lab Tech	Engineering
Mfg & Mgt Engineering	Engineering
Animal & Food Sciences	Life Science
Agriculture	Life Science
Agricultural BioPHYSistry	Life Science
Agricultural Economics	Life Science
Agricultural Educ	Life Science
Agriculture Engr	Life Science
Anatomy & Neurobiology	Life Science
Animal Science	Life Science
Anml Sci & Food & Nutr Science	Life Science
Applied Tech - Ag Engr	Life Science
BioPHYSistry	Life Science
BioPHYSical Science	Life Science
Biology	Life Science
Biostatistics	Life Science
Biological Science	Life Science
Biomedical Technology	Life Science
Botany	Life Science
Biological Sciences	Life Science

Cell Biology	Life Science
Cell & Molec Biology	Life Science
Cellular, Molecular&Biomed Sci	Life Science
Dairy Foods	Life Science
Dental Hygiene	Life Science
Dietetics	Life Science
Dietetics,Nutrition&Food Sci	Life Science
Dairy Technology	Life Science
Ecological Agriculture	Life Science
Forestry	Life Science
Food Systems	Life Science
General Ag Studies	Life Science
Human Nutrition & Foods	Life Science
Lab Animal Tech	Life Science
Microbio & Biophys	Life Science
Medical Microbiology	Life Science
Medical	Life Science
Medical Technology	Life Science
Molecular Genetics	Life Science
Microbiology	Life Science
Medical Laboratory Sciences	Life Science
Medical Laboratory Science	Life Science
Micro & Molec Genetics	Life Science
Neuroscience	Life Science
Nutrition & Food Sciences	Life Science
Nuclear Medicine Technology	Life Science
Nutritional Sciences	Life Science
Pathology	Life Science
Plant Biology	Life Science
Pharmacology	Life Science
Physiology & Biophysics	Life Science
Plant & Soil Science	Life Science
Physical Therapy	Life Science
Radiation Therapy	Life Science
Radiologic Technology	Life Science
Wildlife & Fisheries Biology	Life Science
Wildlife Biology	Life Science
Zoology	Life Science
Mathematics: VMI	Mathematics
Mathematical Sciences	Mathematics
Mathematics	Mathematics
Statistics	Mathematics
Geology	Physical Science
Materials Science	Physical Science

PHYSistry	Physical Science
PhysiComputer Science	Physical Science
Physical Sciences	Physical Science
Chemistry	Physical Science

APPENDIX B

List of UVM's STEM Course Subjects Based on NCES Definition of STEM Fields

Course Subject	Name	frequency	percent
ANNB	ANATOMY & NEUROBIOLOGY	84	0.1
ANPS	ANATOMY/PHYSIOLOGY	1,826	2.16
ASCI	ANIMAL SCIENCE	3,600	4.27
ASTR	ASTRONOMY	721	0.85
BCOR	BIOCORE	4,248	5.03
BIOC	BIOCHEMISTRY	949	1.12
BIOL	BIOLOGY	3,846	4.56
BSCI	BIOLOGICAL SCIENCES	12	0.01
CALS	Agriculture & Life Science	2,037	2.41
CE	CIVIL & ENVIRONMENTAL ENGR	3,321	3.93
CEMS	Engr & Math Sciences	154	0.18
CHEM	CHEMISTRY	8,996	10.66
CIS	COMPUTER INFORMATION SYSTEMS	3	0
CLBI	CELL BIOLOGY	23	0.03
CS	COMPUTER SCIENCE	3,736	4.43
CSYS	COMPLEX SYSTEMS	17	0.02
CTS	CLINICAL&TRANSLATIONAL SCIENCE	2	0
EE	ELECTRICAL ENGINEERING	2,161	2.56
EMGT	ENGINEERING MANAGEMENT	1	0
ENGR	ENGINEERING	979	1.16
ENSC	ENVIRONMENTAL SCIENCES	1,424	1.69
FS	FOOD SYSTEMS	13	0.02
GEOL	GEOLOGY	1,359	1.61
HLTH	HEALTH (HLTH)	2,267	2.69
HSCI	HEALTH SCIENCES (HSCI)	3	0
MAED	MATHEMATICS FOR EDUCATORS	2	0
MATH	MATHEMATICS	10,675	12.65
ME	MECHANICAL ENGINEERING	5,753	6.82
MLRS	MEDICAL LAB & RADIATION SCI	659	0.78
MLS	MEDICAL LABORATORY SCIENCE	560	0.66
MMG	MICR & MOLECULAR GENETICS	1,658	1.96

MPBP	MOLECULAR PHYSIOLOGY & BIOPHYS	42	0.05
NFS	Nutrition and Food Science	6,568	7.78
NMT	NUCLEAR MEDICINE TECHNOLOGY	212	0.25
NSCI	Neuroscience	552	0.65
PATH	PATHOLOGY	104	0.12
PBIO	PLANT BIOLOGY	623	0.74
PHRM	PHARMACOLOGY	4,367	5.17
PHYS	PHYSICS	1,299	1.54
PSS	PLANT & SOIL SCIENCE	2,427	2.88
PSYS	PSYCHOLOGICAL SCIENCE	421	0.5
RADT	RADIATION THERAPY	276	0.33
STAT	STATISTICS	4,753	5.63
WFB	WILDLIFE & FISHERIES BIOLOGY	1,669	1.98

APPENDIX C

STEM Course List by their Codes in Analysis

Course	Name	frequency	percent
AG ADV	Advanced Agriculture	6008	7.19
AG INT	Intermediate Agriculture	3276	3.92
BC 011	Exploring Biology	1146	1.37
BC 012	Exploring Biology	1007	1.21
BC 101	Genetics	840	1.01
BC 102	Ecology and Evolution	553	0.66
BC 103	Molecular and Cell Biology	505	0.6
BC ADV	Advanced Biology	45	0.05
BC INT	Intermediate Biology	110	0.13
BI 001	Principles of Biology	651	0.78
BI 002	Principles of Biology	687	0.82
BI ADV	Advanced Biology	1723	2.06
BI INT	Intermediate Biology	715	0.86
CAL 0I	Calculus I	3198	3.83
CAL II	Calculus I	2512	3.01
CH 023	Outline of General Chemistry	858	1.03
CH 026	Outline of Organic & Biochemistry	629	0.75
CH 031	General Chemistry 1	2545	3.05
CH 032	General Chemistry 2	1475	1.77
CH 141	Organic Chemistry 1	980	1.17
CH 142	Organic Chemistry 2	782	0.94
CH ADV	Advanced Chemistry	1017	1.22
CH INT	Intermediate Chemistry	622	0.74
CS ADV	Advanced Computer Science	1845	2.21
CS INT	Intermediate Computer Science	1826	2.19
EM ADV	Advanced Engineering	7193	8.61
EM INT	Intermediate Engineering	5154	6.17
MA 009	College Algebra	345	0.41
MA 052	Fundamentals of Mathematics	304	0.36
MA 121	Calculus III	892	1.07
MA ADV	Advanced Mathematics	2379	2.85
MA INT	Intermediate Mathematics	901	1.08
NF 043	Fundamentals of Nutrition	1191	1.43
NF 053	Basic Concepts of Foods	355	0.42
NF 063	Obesity: What, Why, What to Do?	294	0.35
NF ADV	Advanced Nutrition & Food Science	3172	3.8

NF INT	Intermediate Nutrition & Food Science	1529	1.83
NH ADV	Advanced Nursing and Health	2557	3.06
NH INT	Intermediate Nursing and Health	1383	1.66
PH 011	Elementary Physics	488	0.58
PH 012	Elementary Physics	386	0.46
PH 021	Introductory Lab I	452	0.54
PH 022	Introductory Lab II	387	0.46
PH 051	Fundamentals of Physics I	188	0.23
PH 152	Fundamentals of Physics II	120	0.14
PH ADV	Advanced Physics	1032	1.24
PH INT	Intermediate Physics	1294	1.55
SC ADV	Advanced Science	7042	8.43
SC INT	Intermediate Science	4243	5.08
ST 111	Elements of Statistics	1179	1.41
ST 141	Basic Statistical Methods 1	1525	1.83
ST ADV	Advanced Statistics	1700	2.04
ST INT	Intermediate Statistics	291	0.35

APPENDIX D

12 Students Clusters with the most Frequent Course Taking Patterns

sequences	Support	Count	sequences	Support	Count
Non-STEM and Switchers			Switchers		
(CAL 0I) → (CAL II)	0.48	357	(CAL 0I) → (SWITCH)	59%	355
(CAL 0I) → (SWITCH)	0.40	301	(CAL 0I) → (CAL II)	52%	314
(CAL 0I) → (NOSTEM)	0.39	289	(SC INT) → (SWITCH)	38%	232
(CAL 0I) → (ST 141)	0.33	247	(CAL II) → (SWITCH)	36%	218
(CAL 0I) → (SC INT)	0.32	240	(CH 031) → (SWITCH)	36%	217
(SC INT) → (SWITCH)	0.28	210	(CAL 0I) → (ST 141)	35%	213
(CAL 0I, CH 031)	0.25	188	(CAL 0I) → (SC INT)	34%	204
(CAL II)-(NOSTEM)	0.25	187	(CAL 0I) → (CAL II) → (SWITCH)	33%	202
(CAL II) → (SWITCH)	0.24	181			
Non-STEM			Engineering		
(NF 043) → (NOSTEM)	0.79	438	(EM INT, EM INT) → (EM ADV, EM INT) → (EM ADV)	0.89	425
(CH 023) → (NOSTEM)	0.73	405	(EM INT) → (EM INT, EM INT) → (EM INT) → (EM ADV)	0.89	424
(NF 043) → (SC INT)	0.72	399	(MA ADV) → (EM INT)	0.89	424
(SC INT) → (SC INT)	0.72	398	(EM INT) → (EM ADV) → (EM ADV) → (EM ADV) → (EM	0.89	423
(SC INT) → (NOSTEM)	0.70	389	(MA 121) → (EM ADV)	0.89	423
(CH 023) → (SC INT)	0.67	369	(MA 121) → (EM ADV, EM ADV)	0.89	423
(NF 043) → (SC INT) → (NOSTEM)	0.66	363	(MA 121) → (EM ADV, EM ADV, EM ADV)	0.89	423
(ST 111) → (NOSTEM)	0.65	361	(MA 121) → (EM ADV, EM ADV, EM ADV) → (EM ADV)	0.89	423
Quitters			Life Science/Food Science		
(CAL 0I) → (QU 100)	66%	315	(NF ADV, NF ADV)	1.00	226
(CH 031) → (QU 100)	58%	279	(NF ADV) → (NF ADV)	1.00	226
(CAL 0I, CH 031)	46%	218	(NF ADV, NF ADV) → (NF ADV)	1.00	225
(CAL 0I, CH 031) → (QU 100)	42%	199	(NF ADV) → (NF ADV, NF ADV)	1.00	225
(BC 011) → (QU 100)	41%	198	(NF INT) → (NF ADV)	1.00	225
(BC 011, CH 031)	38%	180	(NF INT) → (NF ADV, NF ADV)	1.00	225
(CAL 0I) → (CAL II)	37%	179	(AG ADV) → (NF ADV)	0.99	224
Life Science/Agriculture			Math and Computer Science		
(AG ADV, AG ADV) → (AG	1.00	286	(MA ADV) → (MA ADV)	0.67	208

(AG ADV) → (AG ADV)	1.00	286	(MA 121) → (MA ADV)	0.63	198
(AG ADV) → (AG ADV, AG ADV)	0.99	284	(CAL II) → (MA ADV)	0.62	193
(AG ADV, AG ADV) → (AG ADV)	0.98	281	(CAL 0I) → (Cal II)	0.61	190
(AG ADV) → (AG ADV) → (AG ADV)	0.98	281	(CS INT) → (CS ADV)	0.59	183
(AG ADV, AG ADV) → (AG ADV, AG ADV)	0.97	277	(CS INT) → (MA ADV)	0.58	182
(AG ADV) → (AG ADV) → (AG ADV, AG ADV)	0.97	277	(CAL 0I) → (CS INT)	0.58	180
(AG ADV) → (AG ADV, AG ADV) → (AG ADV)	0.96	274	(CAL II) → (CS INT)	0.57	179
(AG ADV) → (AG ADV) → (AG ADV) → (AG ADV)	0.95	271	(CAL II) → (MA 121)	0.56	174
(AG ADV, AG ADV, AG ADV)	0.94	286	(MA ADV) → (MA ADV)	0.67	208
Nursing and Health science			Quitter 2		
(NH ADV, NH ADV)	1.00	110	(EM INT) → (EM INT)	0.73	181
(NH ADV) → (NH ADV)	1.00	110	(CAL 0I, CH 031)	0.71	177
(NH ADV) → (NH ADV, NH ADV)	1.00	110	(CH 031, EM INT)	0.71	176
(NH ADV, NH ADV) → (NH ADV)	0.98	108	(CAL 0I, EM INT)	0.67	168
(NH ADV) → (NH ADV) → (NH ADV)	0.97	107	(CH 031) → (EM INT)	0.65	161
(NH ADV) → (NH ADV) → (NH ADV, NH ADV)	0.97	107	(EM INT, PH INT) →	0.64	160
(NH ADV, NH ADV) → (NH ADV, NH ADV)	0.96	106	(CAL 0I, CH 031, EM INT)	0.62	154
(NH ADV) → (NH ADV, NH ADV) → (NH ADV)	0.96	106	(EM INT) → (PH INT)	0.62	154
(NH ADV, SC ADV)	0.95	105	(CH 031) → (PH INT)	0.61	153
(NH ADV) → (NH ADV, NH ADV)	0.95	104	(EM INT) → (QU 100)	0.59	148
Life Science with Chemistry and Biology			Life Science		
(BC 101) → (BI ADV)	0.90	316	(SC ADV) → (SC ADV)	0.92	544
(CH 032) → (CH 141)	0.89	312	(SC ADV, SC ADV)	0.87	516
(CH 031) → (CH 032)	0.87	306	(SC ADV) → (SC ADV, SC ADV)	0.84	500
(CH 141) → (BI ADV)	0.86	303	(SC ADV) → (SC ADV) → (SC ADV)	0.82	485
(CH 031) → (CH 141)	0.85	299	(CAL 0I) → (SC ADV)	0.75	443
(CH 032) → (BI ADV)	0.85	299	(SC ADV, SC ADV) → (SC ADV)	0.74	441
(CH 031) → (BC 101)	0.85	298	(SC ADV) → (SC ADV) → (SC ADV, SC ADV)	0.73	435
(CH 141) → (CH 142)	0.85	298	(CAL 0I) → (SC ADV) → (SC ADV)	0.70	415